Smooth Risk Functions for Self-Controlled Case Series Models

Thesis

Smooth Risk Functions for Self-Controlled Case Series Models

by

Yonas Ghebremichael Weldeselassie

BA in Statistics and Demography

MSc in Biostatistics

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy in Statistics

Department of Mathematics and Statistics

Faculty of Mathematics, Computing & Technology

The Open University

Walton Hall, Milton Keynes, MK7 6AA

United Kingdom

Date of Submission: 10 March 2014

Date of Award: 19 May 2014

March 2014

Abstract

The self-controlled case series (SCCS) method is commonly used to investigate associations between vaccine exposures and adverse events (side effects). It is an alternative to cohort and case control study designs. It requires information only on cases, individuals who have experienced the adverse event at least once, and automatically controls all fixed confounders that could modify the true association between exposure and adverse event. However, time-varying confounders (age, season) are not automatically controlled.

The SCCS method has parametric and semi-parametric versions in terms of controlling the age effect. The parametric method uses piecewise constant functions with a priori chosen age groups and the semi-parametric method leaves the age effect unspecified. Mis-specification of age groups in the parametric version may lead to biased estimates of the exposure effect, and the semi-parametric approach runs into computational problems when the sample size is moderately large. Moreover, both versions of SCCS represent the time-varying exposures using step functions with pre-determined cut-points. A less prescriptive approach may be beneficial when the shape of the relative risk function associated with exposure is not known a priori, especially when exposure effects can be long-lasting.

This thesis focuses on extending the SCCS method to avoid the aforementioned limitations by modelling the age and exposure effects using flexible smooth functions. Specifically, we used penalised regression splines based on cubic M-splines, which are piecewise polynomials of degree 3. We developed three new extensions: a method that represents only the age effect with splines, a method that uses splines to model only the exposure effect and a non-parametric SCCS method that represents both effects by splines. Simulation studies showed that these new methods outperformed the parametric and semi-parametric methods. The new methods are illustrated using large data sets.

Review of SCCS vaccine studies and directions on how to use the method are also given.

Dedicated to my parents,

sisters, brothers, wife and son

Acknowledgements

First and foremost, I would like express my deepest gratitude to my supervisors Dr. Heather Whitaker and Prof. Paddy Farrington for their inspiring and remarkable guidance throughout the PhD work. I am very grateful to both, for their invaluable supervision, encouragement, kindness and support without which this study would hardly have been completed. Heather and Paddy, I am truly fortunate to have had the chance to work with you. I would also like to express my appreciation to my viva examination panel Prof. Chris Jones, Prof. Frank Critchley and Dr. Irene Peterson for the pleasant discussion we had during the viva.

I extend my gratitude to all members of the statistics group at the Open University for providing an excellent working environment and support in various forms. The financial support from the Open University is greatly acknowledged. Further, I would like to thank my office mates Alexandre and Sofia and colleagues Dr. Angela, Dr. Doyo, Dr. Fadl, Dr. Osvaldo and Tsegay for their help in one way or another. I would also like to acknowledge the Mathematics and Statistics department at Mekelle University, Ethiopia.

I am deeply grateful to my friends Amanuel, Dr. Birhanu, Girma (Woubshet), Hailemariam, Hellen, Miki, Dr. Neguse, Dr. Yoseph, Zenash, and Zerit with their respective families and many others. Your advice, encouragement, support and companion throughout the research have been very important.

My very special thanks go to my parents (Ato G.michael weldeselassie and W/ro Tecle Abraha), my brothers (G.medhin, Berhane, Tesfu, Nahom), my sisters (Askual, Abeba, Semhar), my best friend and brother Desta, my nephews and my inlaws for their never ending love, support, encouragement and prayer. I am also highly indebted to my wonderful wife Saba for her overwhelming love, patience and care and to our son Benhur who has been the source of our happiness and joy in our life. I dedicate this thesis to my family.

And above all I praise the Almighty God for giving me health, strength, courage and the right people in my side to pursue my PhD.

List of Publications

Weldeselassie, Y. G., Whitaker, H. J., and Farrington, C. P. (2011) Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. *Epidemiology and Infection* **139,** 1805–1817.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. (2014). Self controlled case series method with smooth age effect. *Statistics in Medicine* **33(4),** 639 – 649.

Noufaily, A., **Ghebremichael-Weldeselassie, Y.**, Enki, D.G., Garthwaite, P., Andrews, N., Charlett, A., and Farrington, P. (2014) Modeling reporting delays for outbreak detection in infectious disease data. *Journal of Royal Statistical Society Series A*; Accepted.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. (2014). Flexible modelling of vaccine effect in self-controlled case series models. *Journal of Royal Statistical Society Series C*; Submitted.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. (2014). Non-parametric self-controlled case series method. *To be submitted.*

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
CI	Confidence Interval
CV	Cross-Validation
DTP	Diphtheria/Tetanus/Pertusis
Hib	Haemophilus influenza type B
EW	England and Wales
GPRD	General Practice Research Database
ISE	Integrated Squared Error
MISE	Mean Integrated Squared Error
ML	Maximum Likelihood
MMR	Measles, Mumps, Rubella
MSE	Mean Squared Error
RI	Relative Incidence
RMSE	Root Mean Squared Error
SCCS	Self-Controlled Case Series
WCE	Weighted Cumulative Exposure

Chapter 1

Introduction

Although vaccines or other drugs are tested extensively for relatively common adverse events (side effects) in clinical trials before they are licensed for use, not enough people are usually included in such trials to detect adverse reactions that occur only rarely. That is, the randomised double-blind controlled clinical trials used to assess the efficacy of vaccines and drugs before they are licensed are usually insufficiently powered, or too brief, to assess rare but serious side effects or modest increases in the risk of common disease outcomes that have a major population impact in absolute terms (Grosso *et al.*, 2011). Therefore vaccines and drugs used by the wider population need to be constantly investigated for safety.

In addition to assessing the risk of rare events, post-licensure studies also enable the evaluation of safety within groups such as the elderly, those with chronic medical conditions, and pregnant women, who might be deliberately excluded from vaccine or other drug trials. In the context of vaccine safety, by providing accumulating evidence, they can help to maintain the public confidence needed to keep vaccination uptake high enough to prevent disease outbreaks.

Cohort and case-control study designs are commonly used methods to investigate the safety of drugs already on the market. The cohort method compares the risk of a potential adverse effect (outcome event) between individuals who are exposed to the drug of interest and those who are unexposed. This method, although effective, may have a potential problem of confounding variables (Farrington *et al.*, 1996), because the exposed group and unexposed group of individuals could have different characteristics (socio-economic status, underlying health status, gender etc). Confounding variables are variables that are related to both exposure to the drug of interest and the outcome event. These variables, which might be difficult to measure and control, can alter the apparent relationship between the exposure and an outcome event.

The case-control method compares individuals who experienced the outcome event (cases) with individuals who did not experience the event (controls). Controls are usually chosen to be matched to cases on variables like gender, age etc. Case-control studies are less costly and faster to implement than cohort studies. However, as with cohort studies, they suffer from the problem of potential confounding variables that might bias the estimates, and may be associated with difficulties in selecting appropriate controls.

Alternative methods to cohort and case-control methods are study designs that use information only on individuals who have experienced the outcome event at least once (cases). These methods are attractive for three reasons listed in Farrington (2004). First, they can usually be implemented using data extracted from readily available databases such as hospital admission data or other case reporting mechanisms. Second, they can produce results quickly, for example, in response to public concerns or media attention about vaccine or other drug safety. Third, they are usually cheaper to carry out than methods requiring explicit denominators or separate controls.

One of these methods is the self-controlled case series (SCCS) method, or case series method in short, that often combines the power and simplicity of the cohort design and the economy of the case-control method, while eliminating confounding by all time independent variables (variables that do not change their value with time) (Farrington, 1995; Farrington and Whitaker, 2006). It was originally developed, by Farrington (1995), specifically for use in vaccine safety studies, but has since been applied in non-vaccine pharmacoepidemiology and in other areas of epidemiology (Whitaker *et al.*, 2006; Weldeselassie *et al.*, 2011; Grosso *et al.*, 2011).

In the SCCS method, a post-vaccination or duration of drug use risk period (exposure period) is defined a priori, and other times within the period during which each individual is observed (the observation period) constitute the control periods. Then the SCCS method compares the rate of incidence of an event in an exposure period with the rate of incidence in the control periods, when an individual is not exposed. The comparison is within individuals. The incidence rate in the control period is the baseline incidence rate; this is not estimated in the SCCS method. The estimated measure of the relationship between exposure and outcome event is a relative incidence. Because the comparison is made within an individual's observation period, the method is self-matched; hence, all measured and unmeasured age-independent confounding variables, such as socio-economic status, birth weight, location, severity of underlying disease, gender, etc., which act multiplicatively on the baseline incidence rate, are automatically controlled. However, time-varying confounders such as age and season are not automatically controlled for, but as with cohort methods they can be allowed for explicitly in the model (Farrington, 2004).

Careful control of age effects is particularly important in the study of paediatric vaccines and neurological events, such as febrile convulsions. The incidence of such events is

highly age-dependent in the first two years of life, which is precisely the age at which many routine vaccinations take place. Partly for this reason, potential associations between vaccination and neurological events have been studied intensively over several decades. These studies have used a broad range of methods, including SCCS (Farrington *et al.*, 1995; Barlow *et al.*, 2001; Huang *et al.*, 2010; Miller *et al.*, 1981). Similarly the effect of exposure, the main focus of interest, should be modelled appropriately and carefully.

In its original form, the case series model took the multiplicative effect of age on the baseline incidence rate into account by dividing age into selected groups, with the age effect being represented by a piecewise constant step function. That is, the age effect is taken to be constant over the chosen age groups. We refer to this as the parametric version of the case series method. Its limitation is that it can be sensitive to mis-specification of the a priori selected age groups, which may lead to biased estimates of the association between exposure and outcome event . Another version of the SCCS method, in terms of modelling the age effect, is the semi-parametric SCCS (Farrington and Whitaker, 2006). In this method, the function that represents the age effect is not specified a priori, hence avoiding the limitation of the standard (parametric) SCCS method. However, as the number of cases in the study increases, the number of parameters to be estimated increases, which leads to computational problems (Farrington and Whitaker, 2006). In both the parametric and the semi-parametric versions of the SCCS method, the effect of age is represented by a step function; in the parametric version these age groups are chosen a priori, whereas in the semi-parametric version they are determined by the data.

The effect of exposure, in both versions, is modelled as a step function based on groups chosen a priori. Similar to the age effect, the use of step functions to model the exposure effect might have limitations: a poor choice of cut-points may be associated with cut-point

bias and misclassification (Altman, 1991; Greenland, 1995b).

In this thesis, to avoid these limitations, we replace the step functions in the parametric and semi-parametric SCCS methods by smooth functions that are based on M-splines. M-splines are piecewise polynomial functions connected at points known as knots and their linear combination is known as a spline function. The likelihood function of the SCCS method, which may be derived from a cohort method by conditioning on the total number of events experienced by each individual, contains an integral in its denominator. The use of M-splines to represent the age and exposure effects not only removes the limitations of step functions, but also avoids the numerical integration of the integral in the likelihood function, because the integral of an M-spline can be expressed in terms of another spline known as an I-spline. The other advantage of using M-splines is that they are positive functions and therefore can be used to approximate a non-negative function by constraining their coefficients to be non-negative. In SCCS, the functions that represent age and exposure effects should be non-negative functions as they are relative effects.

1.1 Aims and Objectives

The main objective of the thesis is to develop new extensions to the self-controlled case series method in order to circumvent the limitations associated with the use of step functions in the standard and semi-parametric versions of the method. But first, to set the scene, a review of how the SCCS method has been used in vaccine studies is undertaken, and clear directions on how it should be used are given. Therefore the aims and objectives are:

- to investigate the limitations of the standard and semi-parametric SCCS methods using a simulation study;

- To review how the self-controlled case series method has been applied in vaccine studies, clarify misconceptions about the method and present some recommendations on how it should be used, with the emphasis on promoting good practice;
- to represent the age effect with smooth functions as a linear combination of M-spline functions while representing exposure effect by a step function;
- to model the exposure-related relative incidence function using a linear combination of M-splines and use a piecewise constant function for the age effect; and
- to represent both the age and exposure effects using linear combinations of M-splines at the same time.

1.2 Thesis Outline

The thesis begins with a description of the self-controlled case series method and a derivation of the likelihood functions of the standard and semi-parametric versions of the method in **Chapter** 2. In addition, **Chapter** 2 presents a simulation study conducted to investigate the limitations of the standard and semi-parametric SCCS methods.

Chapter 3 presents a critical review of vaccine studies that made use of the SCCS method between 1995 and beginning of 2014. This review includes discussion on: how the studies described their data and accuracy of the data, how observation periods and risk periods were chosen, how potential confounders were handled, potential sources of biases, comparison of SCCS results with other statistical methods, some methodological issues, power and sample size issues and how sensitivity analyses were done. Also, recommendations on how the SCCS method should be used and reported are given.

Chapter 4 introduces some of the smooth functions which could be used in modelling

age and exposure effects to avoid the limitations investigated in **Chapter** 2. The functions discussed in this chapter are polynomial functions, fractional polynomials, truncated power functions, B-splines, M-splines and I-splines.

Chapters 5, 6 and 7 describe the extensions made to the standard SCCS method by replacing the step functions which represent age and exposure effects by smooth functions. In **Chapter** 5 the standard SCCS method is extended by representing the age-related relative incidence function as a linear combination of cubic M-splines (piecewise cubic polynomial functions) and the cumulative age-specific relative incidence by a linear combination of I-splines. The use of polynomial and fractional polynomials in the context of the SCCS method are also described in this chapter. To use M-splines to represent a function, it is first necessary to determine the number and position of the knots used to make them. Smoothing splines are spline functions where the knots are placed at data points: so prior knots do not need to be selected. This however greatly increases the computational burden, because the number of parameters to be estimated is about equal to the number of observations. Instead we use penalised regression splines, for which the number of knots is less than the number of observations and the knots need to be determined a priori. Selecting too small a number of knots under-fits the function and too large a number of knots over-fits it. So we use a large number of knots and introduce a penalty term to the log-likelihood function to control the roughness of the function. **Chapter** 5 presents the derivation of a penalised log-likelihood function for the spline-based SCCS method. An approximate cross validation method used to choose a smoothing parameter that controls the tradeoff between roughness and fit is also derived in this chapter. Finally, **Chapter** 5 presents a simulation study conducted to evaluate the performance of the new method and its application to investigate a potential association between pedi-

atric vaccines and febrile convulsion. In this chapter the exposure effect is represented by a step function.

Chapter 6 presents a different extension of the standard SCCS method that models the exposure-related relative incidence function as a linear combination of cubic M-splines while the age effect is represented by a step function. Similar to **Chapter** 5, a penalised log-likelihood function is used. A simulation study to investigate the performance of this method relative to the standard method, and applications to data on pediatric vaccines and febrile convulsions, and thiazolidinedione (a class of medicines used to treat type 2 diabetes) use and fractures are also presented.

In **Chapters** 5 and 6 step functions are still used to model one of either the exposure or the age effect. In **Chapter** 7, the works of **Chapters** 5 and 6 are combined in developing a non-parametric SCCS method, where both age and exposure effects are approximated by linear combinations of cubic M-splines. In this chapter we define the first, second and third integrals of an I-spline based on the definition for the integral of an M-spline given by Ramsay (1988). We also define the integral for the product of two spline functions expressed as linear combinations of cubic M-splines. **Chapter** 7 also presents a simulation study that evaluates the performance of the non-parametric SCCS method.

Finally, conclusions and possible future research are presented in **Chapter** 8.

Chapter 2

The Self-Controlled Case Series Method

The self-controlled case series method, which uses only information from cases, that is, individuals with an adverse event, was developed specifically for use in vaccine safety studies, but has since been applied in non-vaccine pharmacoepidemiology and in other areas of epidemiology (Whitaker *et al.*, 2006). It automatically controls all age-independent multiplicative confounders, while allowing for an age-dependent baseline incidence. The method has two versions based on the way it handles the effect of age-dependent confounding variables: (1) the standard method, which models the effect of age using a parametric step function and (2) the semi-parametric method: that controls for age non-parametrically. In this chapter we describe how the two versions of the method work and derive their likelihood functions. And limitations of the SCCS models, which led to the extensions developed in this thesis, are investigated using a simulation study.

In Section 2.1, how the SCCS method works is described, followed by the derivation of the likelihood function for the standard version of the method and a general likelihood

function in Section 2.1.1. Section 2.2 deals with the semi-parametric version of the method and how to fit the model. Finally, the limitations of the standard and semi-parametric SCCS models are given in Section 2.3 followed by a discussion in Section 2.4.

2.1 The Standard Case Series Method

The self-controlled case series method, in its standard framework, was developed to estimate the relative incidence of an acute event in a pre-defined post-vaccination risk period (Farrington, 1995). The relative incidence is a ratio of the incidence rate in a predefined post-exposure risk period relative to other times (control periods) within a defined period during which individuals are observed (the observation period). It is a conditional, retrospective, risk-interval cohort method and is applied as follows.

An overall study time-window, usually defined by age and calendar time boundaries (and also, sometimes, in terms of vaccination date), is chosen, ideally such that the chance that individuals experience both risk and control periods is maximised. The observation period, in particular, must be defined so that, had an event occurred at any point within it, the case would have been ascertained. Then, all or a random sample of individuals with at least one event (independent recurrences are permitted) within this study time-window are identified: these are the cases. The study time-window also determines individual observation periods for each case, namely the time spent by each individual within the study time-window (the observation periods generally differ between individuals).

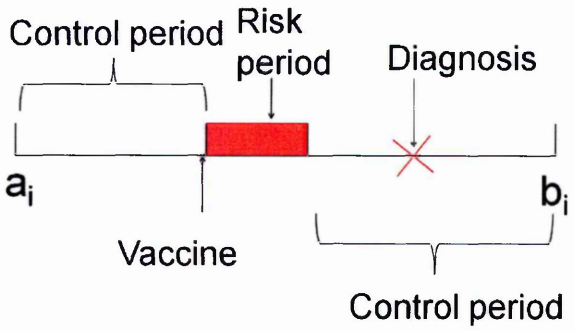

Figure 2.1: *Self-controlled case series setup; where a_i and b_i are ages at the start and end of observation period for individual i respectively*

Next, the vaccination or exposure to other drugs histories of the cases are collected. As in other epidemiological designs, ascertainment of cases must be independent of vaccination or exposure histories, dependence of exposure history on case ascertainment may lead to biased relative incidence estimates. A clear description of how the data were obtained is therefore important in order for the reader to be able to assess any possible dependence. The vaccination dates of each case are used to define one or more risk periods, during which individuals are hypothesized to be at increased (or reduced) risk of the event of interest after (or, for reasons to be discussed later, before) vaccination. Risk periods are defined in terms of time since vaccination (with, preferably, a stated convention to describe the day of vaccination, for example day zero) or duration of exposure to drugs other than vaccines, which are not point exposures. A rigorous report of these choices provides confidence that care was taken in the analysis, and enables the reader, in theory at least, to reconstruct the study exactly. In the context of point exposures like vaccine, the choice of risk periods should be made a priori and its rationale explained. Typically, the choice will be motivated by reference to previous studies or hypotheses, by biologically plausible mechanisms or by expert opinion. All other time within an individual's observation pe-

riod, that does not fall within a risk period, is included in that individual's control period, which forms the study baseline (see Figure 2.1). Figure 2.1 shows an observation period $(a_i, b_i]$ for individual i divided into control and exposure groups. Diagnosis is the age at which the event of interest occurs (it can be anywhere within the observation period and an individual can experience more than one event provided that they are independent) and a_i and b_i are ages at the start and end of the observation period respectively.

Justification for using only cases stems from the analytical strategy, which conditions on the number of events each individual experiences within the observation period: this number is regarded as a fixed quantity. A consequence is that non-cases contribute no information, and therefore need not be sampled. Estimation of parameters in the SCCS method is achieved by fitting a conditional Poisson regression model (it is essential that it should be a conditional model, in order to justify sampling only cases). The parameter of interest is the relative incidence, that is, the incidence in a risk period relative to the control or baseline periods. A further consequence of the conditioning is that the analysis is within-individuals, and, as a result, in the SCCS method all fixed confounding factors, known and unknown, are controlled for implicitly. Temporal confounding factors, such as age can be accounted for by subdividing each individual's observation period into age categories, which are modelled explicitly.

2.1.1 Derivation of SCCS Likelihood

In this section we derive the likelihood function used to estimate the parameter of interest (relative incidence) in the standard SCCS method, where the effect of age is taken into account by dividing the observation period into age groups. We also derive a general likelihood function for the SCCS model where the age effect can be represented

by a variety of functions.

Standard SCCS Likelihood Function

The assumptions made in deriving the likelihood function are: Assumption (1) that individuals experience events in a non-homogenous Poisson process; Assumption (2) that age-dependent exposures experienced by individuals are exogenous, so exposures are independent of prior events, and Assumption (3) that censoring of individuals at the end of the observation period occurs completely at random, i.e the occurrence of the event of interest must not censor or affect the observation period Farrington (1995); Farrington and Whitaker (2006); Whitaker *et al.* (2006); Weldeselassie *et al.* (2011). Discussion about deviations from these assumptions is given in Chapter 3.

Let $(a_i, b_i]$ be the observation period for individual $i = 1, 2, ...N$, often determined by a combination of calendar time and age constraints. And let individual i be exposed at age c_i and the risk period be $(c_i, d_i]$ so that k is an indicator of exposure, $k = 1$ in the risk period and $k = 0$ in the control periods as shown in Figure 2.2. More than one exposure periods are possible. In order to control for age, if for example the number of age groups is 2, the observation period in Figure 2.2 is further divided into two segments as $j = 1$ and $j = 0$.

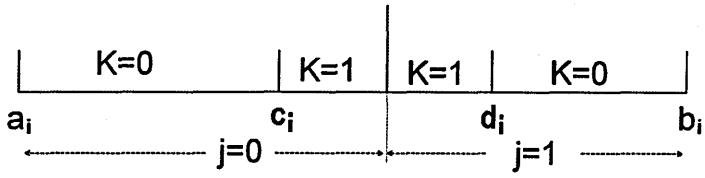

Figure 2.2: *Self-controlled case series model with two age and two exposure groups*

In Figure 2.2, for this particular individual i, the observation period is divided into 4 intervals. Let the length of interval (i, j, k) be denoted by e_{ijk} and the number of events

experienced in the interval be denoted by n_{ijk}. The disease incidence rate λ_{ijk} is assumed to be constant within an interval. Denote the baseline incidence rate of individual i in the age group $j = 0$ and exposure group $k = 0$ by $\varphi \exp(\gamma_i)$ and let $\exp(\alpha_1)$ and $\exp(\beta_1)$ denote relative incidences associated with age group $j = 1$ and risk period $k = 1$ respectively, relative to age group $j = 0$ and exposure group $k = 0$ (with $\alpha_0 = \beta_0 = 0$) , where φ and $\exp(\gamma_i)$ are age-independent fixed and random individual effects that may depend on covariates that do not vary over the period $(a_i, b_i]$.

In the SCCS method, exposure, age and other variables are assumed to have a multiplicative effect on the baseline rate of incidence. And since events are assumed to arise in a non-homogenous Poisson process, constant within an interval, the Poisson rate in the first interval within the observation period of individual i, shown in Figure 2.2, is:

$$\begin{aligned} e_{i00}\lambda_{i00} &= e_{i00}\varphi \exp(\gamma_i) \exp(\alpha_0) \exp(\beta_0) \\ &= e_{i00}\varphi \exp(\gamma_i), \end{aligned}$$

as $\alpha_0 = \beta_0 = 0$. Similarly the rates in the second, third and fourth intervals are respectively

$$\begin{aligned} e_{i01}\lambda_{i01} &= e_{i01}\varphi \exp(\gamma_i) \exp(\beta_1), \\ e_{i11}\lambda_{i11} &= e_{i11}\varphi \exp(\gamma_i) \exp(\alpha_1) \exp(\beta_1), \\ e_{i10}\lambda_{i10} &= e_{i10}\varphi \exp(\gamma_i) \exp(\alpha_1). \end{aligned}$$

In a case series analysis an individual is included in a sample if at least one event occurred in his/her observation period. Hence we condition on the number of events an individual experienced, in deriving the likelihood contribution of the individual. This leads to a multinomial distribution. Then the likelihood contribution of individual i according

to the multinomial distribution, 4-nomial in this example, is:

$$L(\mathbf{p}_i|n_i, y_i) = \frac{n_i!}{y_{i1}!y_{i2}!y_{i3}!y_{i4}!} p_{i1}^{y_{i1}} p_{i2}^{y_{i2}} p_{i3}^{y_{i3}} p_{i4}^{y_{i4}} \tag{2.1}$$

where y_{i1}, y_{i2}, y_{i3} and y_{i4} are the numbers of events (the values of n_{ijk}) in the intervals $1, 2, 3$ and 4 respectively of individual i within the observation period. $n_i = \sum n_{i..} = \sum_j \sum_k n_{ijk}$ is the total number of events experienced by individual i. p_{i1}, p_{i2}, p_{i3} and p_{i4} are probabilities of an event to a occur in a corresponding interval. Since the events occur as a Poisson process and are independent, the probabilities can be found from the following property. If $Y_{i1}, Y_{i2}, \ldots, Y_{is}$ are independent Poisson random variables with parameters $\mu_1, \mu_2, \ldots, \mu_s$ then

$$Y_{if} | \sum_{l=1}^{s} Y_{il} \sim Binom\left(\sum_{l=1}^{s} Y_{il}, \frac{\mu_f}{\sum_{l=1}^{s} \mu_l} \right), \tag{2.2}$$

$$Y_{if}|n_i \sim Binom\left(n_i, \frac{\mu_f}{\sum_{l=1}^{s} \mu_l} \right). \tag{2.3}$$

Therefore, the expression for the parameters (probabilities) in the multinomial likelihood function is $\frac{\mu_f}{\sum_{l=1}^{s} \mu_l}$. For example

$$\begin{aligned} p_{i1} &= \frac{e_{i00}\varphi \exp(\gamma_i)}{\varphi \exp(\gamma_i)\left(e_{i00} + e_{i01}\exp(\beta_1) + e_{i11}\exp(\alpha_1 + \beta_1) + e_{i10}\exp(\alpha_1)\right)} \\ &= \frac{e_{i00}}{e_{i00} + e_{i01}\exp(\beta_1) + e_{i11}\exp(\alpha_1 + \beta_1) + e_{i10}\exp(\alpha_1)} \end{aligned}$$

and in general

$$p_{ijk} = \frac{e_{ijk}\exp(\alpha_j)\exp(\beta_s)}{\sum_{rs} e_{irk}\exp(\alpha_r)\exp(\beta_s)}.$$

The individual effect $\exp(\gamma_i)$ and the baseline incidence φ cancel out, which implies that the SCCS method implicitly controls for all measured and unmeasured fixed confounding variables. Therefore, the likelihood contribution of individual i ignoring the constant in the multinomial likelihood since it does not depend on the parameters of

interest, is

$$L_i(\alpha,\beta) = \prod_{jk}\left[\frac{e_{ijk}\exp(\alpha_j+\beta_k)}{\sum_{rs}e_{irs}\exp(\alpha_r+\beta_s)}\right]^{n_{ijk}}.$$

As individuals are independent the likelihood for all individuals is

$$L(\alpha,\beta) = \prod_{i=1}^{N}\prod_{jk}\left[\frac{e_{ijk}\exp(\alpha_j+\beta_k)}{\sum_{rs}e_{irs}\exp(\alpha_r+\beta_s)}\right]^{n_{ijk}}$$

and the log-likelihood function is

$$l(\alpha,\beta) = \sum_{i=1}^{N}\sum_{jk} n_{ijk} log\left[\frac{e_{ijk}\exp(\alpha_j+\beta_k)}{\sum_{rs}e_{irs}\exp(\alpha_r+\beta_s)}\right]. \tag{2.4}$$

The General Likelihood Function

As in the piecewise constant case, the general likelihood function of a case series method may be derived from a cohort likelihood by conditioning on the number of events each individual experiences and on the exposure history over the time period an individual is observed. Suppose that individual i in a dataset is observed in a period $(a_i, b_i]$ where $i = 1,\ldots,N$ and experiences events $t_{i1}, t_{i2}, \ldots, t_{in_i}$. Within this observation period individuals experience exposures from different risks which can change the probability to experience an event. Denote the number of events that an individual experiences in the interval $(a_i, t]$ by $N_i(t)$. As in Farrington and Whitaker (2006), let $x_i(t)$ represent the vector of exposures that individual i experiences at age t within the observation period. If there is one exposure, letting β to denote the effect of exposure $x_i(t)\beta$ or using a different parametrisation $\exp(x_i(t)\beta)$ represents exposure related relative incidence function (see below in this section). Again let x_i^t be the exposure history of individual i up to age t, that is, the function $x_i^t = \{x_i(s) : s \leq t\}$. Let $x_i \equiv x_i^{b_i}$ be the exposure history of individual i up to the end of their observation period. Letting the intensity process (hazard function) by which events arise be denoted by $\lambda_i(t|x_i^t)$, we have a probability density function given by

$\lambda_i(t|x_i^t)S(t|x_i^t)$. Here $S(t|x_i^t)$ is a survival function which can be obtained from the hazard function as $S(t|x_i^t) = \exp\left\{-\int_{a_i}^{t} \lambda_i(u|x_i^t)du\right\}$.

Since events within the observation period of an individual are assumed to be independent, the unconditional likelihood that individual i experiences n_i events that arise with intensity process $\lambda_i(t|x_i^t)$ at times t_{ij}, $j = 1, 2, ..., n_i$ is

$$L_i^u = \prod_{j=1}^{n_i} \lambda_i(t_{ij}|x_i^{t_{ij}}) \exp\left\{-\int_{a_i}^{b_i} \lambda_i(t|x_i^t)dt\right\}. \tag{2.5}$$

Assumption (2) (in page 13) implies that the event rate at age t, given the exposure history to age t, is equal to the event rate at age t, given the exposure history over the entire observation period, i.e. $\lambda_i(t|x_i^t) = \lambda_i(t|x_i)$. Thus, conditioning on the total number of events an individual i experiences in their observation period and on the exposure history x_i does not affect $\lambda_i(t|x_i)$. This is the key assumption of the SCCS method, departures from which are discussed in detail in Farrington *et al.* (2009). Departures from assumption (3) are discussed in Farrington *et al.* (2011).

Now to find the conditional likelihood that an individual i experiences the events t_{ij}, $j = 1, 2, ..., n_i$ conditional on the total number of events experienced by the end of the observation period, we need to have the expression for the probability that the count of events is n_i. In SCCS method events are assumed to occur in a non-homogeneous Poisson process with intensity $\lambda_i(t|x_i)$, therefore the total count $N_i(b_i)$ is a Poisson random variable with mean $\int_{a_i}^{b_i} \lambda_i(t|x_i)dt$). This leads to the probability

$$P(N_i(b_i) = n_i) = \frac{\left\{\int_{a_i}^{b_i} \lambda_i(t|x_i)dt\right\}^{n_i}}{n_i!} \exp\left\{-\int_{a_i}^{b_i} \lambda_i(t|x_i)dt\right\} \tag{2.6}$$

Therefore, from Equations (2.5) and (2.6) the conditional likelihood contribution of

individual i given n_i events occurred at times t_{ij}, $j = 1, 2, ..., n_i$ is obtained as

$$\begin{aligned} L_i^c &= \prod_{j=1}^{n_i} \frac{\lambda_i(t_{ij}|x_i)}{\int_{a_i}^{b_i} \lambda_i(t|x_i)dt} \\ &= \frac{\prod_{j=1}^{n_i} \lambda_i(t_{ij}|x_i)}{\left\{\int_{a_i}^{b_i} \lambda_i(t|x_i)dt\right\}^{n_i}}. \end{aligned} \tag{2.7}$$

From this case series likelihood it can be seen that if individual i has no events, $n_i = 0$, then $L_i^c = 1$, implying that only individuals with at least one event in their observation period contribute to the likelihood. Hence, the case series method needs information only on cases.

The most convenient way of parameterizing the incidence $\lambda_i(t|x_i)$ is according to the proportional incidence model

$$\begin{aligned} \lambda_i(t|x_i) &= \lambda_0(t) \exp\left\{\gamma_i + x_i(t)^T\boldsymbol{\beta}\right\} \\ &= \varphi\psi(t) \exp\left\{\gamma_i + x_i(t)^T\boldsymbol{\beta}\right\}, \end{aligned} \tag{2.8}$$

where $\lambda_0(t) = \varphi\psi(t)$ is the baseline incidence at age t (to be discussed next), γ_i is as defined above, and $\boldsymbol{\beta}$ is a vector of the log-relative incidences that measure the association between exposures and event of interest. Then, combining Equations (2.7) and (2.8), the conditional likelihood contribution of individual i is

$$L_i^c = \prod_{j=1}^{n_i} \frac{\psi(t_{ij}) \exp\left\{x_i(t_{ij})^T\boldsymbol{\beta}\right\}}{\int_{a_i}^{b_i} \psi(t) \exp\left\{x_i(t)^T\boldsymbol{\beta}\right\} dt}.$$

Since individuals are independent the conditional likelihood of all individuals is given as

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\psi(t_{ij}) \exp\left\{x_i(t_{ij})^T\boldsymbol{\beta}\right\}}{\int_{a_i}^{b_i} \psi(t) \exp\left\{x_i(t)^T\boldsymbol{\beta}\right\} dt}. \tag{2.9}$$

The terms φ and $\exp(\gamma_i)$ cancel out and hence all fixed covariates that act multiplicatively on the baseline incidence are automatically controlled for. This is because the total number of events experienced by individual i, n_i, is a sufficient statistic to estimate the

individual effects, $\exp(\gamma_i)$, which leads to the removal of $\exp(\gamma_i)$ from the conditional likelihood.

In the standard SCCS method $\psi(t)$, which is the age-specific relative incidence function, is represented using a step function, so replacing $\psi(t)$ and the exposure effect in 2.9 by step functions yields the log-likelihood function 2.4. In the semi-parametric version of SCCS (Section 2.2) $\psi(t)$ is left unspecified: the cumulative baseline relative incidence function is a step function with steps at the distinct event times.

2.1.2 Fitting the Standard Model

The log-likelihood function of the standard self-controlled case series method can either be maximised directly or fitted using as an associated Poisson regression model with log link function. To fit the associated conditional Poisson model, the data should be formatted such that there is one line for each interval within the observation period (see Figure 2.2). The number of events in an interval n_{ijk} is used as a response variable in the model and log of interval lengths are included as an offset. Factors for age group, exposure group and individual are also listed for each of the intervals. The model fitted is

$$\begin{aligned} n_{ijk} &\sim \text{Poisson}(\lambda_{ijk} e_{ijk}) \\ \log(\lambda_{ijk}) &= \varphi_i + \alpha_j + \beta_k, \end{aligned} \tag{2.10}$$

where φ_i is an individual effect included to guarantee that the fitted marginal totals equal the observed values. As the φ_i are nuisance parameters, it is convenient to fit the model as a conditional fixed effects Poisson regression model.

2.2 Semi-Parametric SCCS

In the standard self-controlled case series method, the age-specific relative incidence is defined to be a step function and fitted by a priori choosing age groups over which the incidence is believed to be roughly constant. An extension of the standard model is a semi-parametric model in which the age-specific relative incidence is left unspecified except that it is non-negative and bounded (Farrington and Whitaker, 2006). The cumulative age-specific relative incidence function $\Psi(t) = \int_a^t \psi(s)ds$ is estimated non-parametrically from non-decreasing step functions. The likelihood function of the semi-parametric method can be derived in a similar way to the standard SCCS method. Let the set of distinct event ages t_{ij} of all cases be denoted by S. Assume that there are M distinct event ages sorted in an increasing order $s_1, ..., s_M$. Let the step function that represents the cumulative age-specific relative incidence function be constant outside S and have jumps of height $\Delta\Psi(t)$, for $t \in S$, where $\Delta\Psi(s_r) = \exp(\alpha_r)$, $r = 1, ..., M$, and without loss of generality let $\alpha_1 = 0$. Define a weight for each individual i and each $s_r \in S$ as $w_{ir} = I_{(a_i,bi]}(s_r)$, where $I_{(a_i,bi]}$ is an indicator which takes a value 1 if s_r is within the observation period of individual i and 0 otherwise. The weights assigned ensure that only event days within the observation period of an individual contribute to his/her likelihood and that the jumps in the cumulative function are at the event ages. Let α_{ij} be the value of α_r corresponding to t_{ij} (the j^{th} event of individual i). The cumulative age-specific relative incidence curve of the semi-parametric model is presented in Figure 2.3.

Figure 2.3: *Representation of age effect in the semi-parametric self-controlled case series method*

Figure 2.3 shows that the cumulative age-specific relative incidence is constant between two event ages and jumps at each of the event ages.

2.2.1 Semi-Parametric SCCS Likelihood

The likelihood function for the semi-parametric model can be derived in a similar way to the standard SCCS by conditioning on the total number of events an individual experiences in their observation period. It can also be derived from the general likelihood function of SCCS (2.9) by replacing the effect of exposure and the cumulative age-specific relative incidence ($\Psi(t) = \int_a^t \psi(s)ds$) with step functions. That is, in the numerator we have $\psi(t_{ij}) = \exp(\alpha_{ij})$ and in the denominator $\int_{a_i}^{b_i} \psi(t) \exp\left\{x_i(t)^T\beta\right\} dt$ is replaced by $\sum_{r=1}^{M} w_{ir} exp(\alpha_r + x_i(s_r)^T\beta)$, which leads to the likelihood contribution of individual i

$$L_i = \prod_{j=1}^{n_i} \frac{exp(\alpha_{ij} + x_i(t_{ij})^T\boldsymbol{\beta})}{\sum_{r=1}^{M} w_{ir} \exp(\alpha_r + x_i(s_r)^T\boldsymbol{\beta})},$$

where $x_i(t_{ij})$, as defined before, represents the exposures. Once again, the individual effects φ and $\exp(\gamma_i)$ cancel out and lead to automatic control of age-independent confounding covariates. The independence of individuals gives the SCCS semi-parametric

likelihood function (L_s) for all individuals

$$L_S = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\exp(\alpha_{ij} + x_i(t_{ij})^T \beta)}{\sum_{r=1}^{M} w_{ir} \exp(\alpha_r + x_i(s_r)^T \beta)}. \tag{2.11}$$

The function $\Psi(t)$ represents a relative effect, and so is not identifiable without some further constraint. In the semi-parametric SCCS method (Farrington and Whitaker, 2006), the constraint $\Psi(a) = 1$ is used, so that φ is the baseline incidence at a.

2.2.2 Fitting the Semi-Parametric Model

Fitting the semi-parametric model is similar to fitting a standard model with unit age groups, where the unit is the smallest separation between successive event times, i.e $min\{s_{r+1} - s_r, r = 1, \ldots, M-1\}$. To fit the model, the data need to be expanded such that each individual has a row of information for each of the distinct ages at event in the data set that falls within their observation period. Information on the number of events experienced by individual i, n_{ir} (which is 0 or 1 since each interval has length of only one time unit), exposure status, $x_i(t_{ij})$, and weight, w_{ir}, at age s_r are included. As an example, consider a study of 6 cases with five distinct event ages 20, 75, 143, 160 and 200 (two cases experienced their events at the same age). Let individual 1 have an observation period (10, 180], exposure period (135, 150] and their event occurred at age 75.

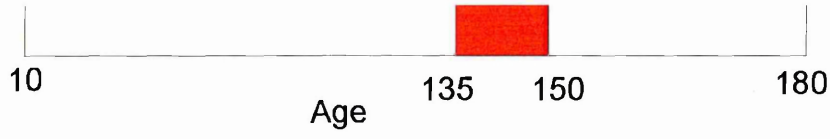

Figure 2.4: *An individual with one exposure period, four of the event ages falling within the observation period*

One of the event ages, 200, is outside the observation period of individual 1, hence

its weight is zero. Therefore, the expanded data for individual 1 will have four rows with non-zero weight as presented in Table 2.1.

Table 2.1: *Reformatted data for one case, used to fit the semi-parametric model*

Id	Events(n_{ir})	Risk group(x_i)	Age group	Weight(w_{ir})
1	0	0	0	1
1	1	0	1	1
1	0	1	2	1
1	0	0	3	1
1	0	0	4	0

Once the data are reformatted the conditional Poisson model is fitted to estimate the parameters of interest. The model has weights w_{ir}, response variable n_{ir} with mean λ_{ir} and a log link function

$$log(\lambda_{ir}) = \gamma_i + \alpha_r + x_i(s_r)\beta$$

where γ_i is an individual effect included to constrain the total number of events experienced by each individual to their observed values.

2.3 Limitations of SCCS

This section explores the limitations of the standard and semi-parametric versions of the self-controlled case series method. In the standard SCCS method the age groups should be specified a priori as described in Section 2.1.1. We, therefore, investigate the sensitivity of the parameter estimates related to exposure, in the standard SCCS method, to mis-specification of age groups. In the semi-parametric method the number of parameters that need to be estimated increases with the sample size because the fitting procedure using standard software requires the data to be expanded to size of order N^2, where N is the number of cases. The age effect is represented by a vector of parameters whose

dimension is of order N. When N is moderately large, this may lead to computational problems. There may also be a loss of efficiency in estimation. We investigate these using simulation studies.

2.3.1 Limitation of the Standard SCCS

A major limitation of the standard SCCS method is that estimates of the exposure effect may be biased if the a priori chosen age groups are misspecified (Farrington and Whitaker, 2006). To investigate this limitation we conducted a simulation study. In the simulation study we selected the beginning and end of the observation period for all individuals to be 0 and 730 days respectively. The ages at exposure, c_i had an exponential distribution, and were generated from an exponential distribution with a rate of 0.01. We took the risk period to be 50 days post exposure. Three values of the true exposure-related relative incidence were investigated: 1, 2, and 5.

15 age groups were used in simulating the data, with cut points at every 50 days between 0 and 730 with the last age group having a length of 30 days and the true age-specific relative incidence values were 1, 2, 3, 5, 6, 7, 9, 10, 11, 13, 15, 16, 18, 19 and 20. Three different scenarios of sample size (number of cases), 50, 100 and 200, were considered. For each scenario 10,000 data sets were generated, (for more on how data are generated in the SCCS method see Section 5.3 of Chapter 5).

The 10,000 simulated data sets were then analysed using the SCCS method without any age effect, the standard SCCS with misspecified age groups (two age groups separated at age of 350 days) and the standard SCCS method with the 15 correctly specified age groups used in generating the data sets. Results of these analyses are presented in Table 2.2. We computed bias and standard error of the bias for each scenario. The biases

were calculated as:

$$\begin{aligned} \text{Bias} &= \text{median of the 10,000 estimated exposure-related log relative incidences} \\ &- \text{the true log relative incidence.} \end{aligned} \tag{2.12}$$

Table 2.2: *Simulation study results of investigating the effect of age groups' mis-specification in the standard SCCS method. Bias of the exposure-related log relative incidence and their standard errors (SE) are presented*

Number of Cases	No age effect included	Misspecified	Correctly specified
	Bias(SE)	Bias(SE)	Bias(SE)
	True Relative incidence = 1		
50	0.516(0.020)	0.168(0.020)	-0.024 (0.020)
100	0.657 (0.004)	0.201(0.004)	-0.012 (0.004)
200	0.599 (0.003)	0.177 (0.002)	-0.005 (0.003)
	True Relative incidence = 2		
50	0.527 (0.004)	0.196 (0.005)	-0.008 (0.005)
100	0.589 (0.003)	0.178 (0.003)	-0.004 (0.003)
200	0.590 (0.002)	0.183 (0.002)	0.002 (0.002)
	True Relative incidence = 5		
50	0.554 (0.003)	0.196 (0.003)	0.035 (0.004)
100	0.585(0.002)	0.176 (0.002)	0.022 (0.003)
200	0.593 (0.002)	0.189 (0.002)	0.007 (0.002)

We used the median in calculating the bias because it is possible for all event ages to occur in the risk period only, or in the control period only, resulting in an undefined expected value of the estimated relative incidence. The standard errors of the biases were also calculated by trimming the unbounded estimates. The standard errors were

calculated as

$$SE = \frac{SD(\hat{\beta} - \beta)}{\sqrt{N^*}}$$

where $\hat{\beta}$ is the estimated exposure-related log-relative incidence, β is the true log-relative incidence, N^* is the number of data sets with bounded estimates and SD is standard deviation of biases of each estimate.

The results in Table 2.2 show that when age is misspecified or ignored from the standard SCCS analysis the estimated exposure-related relative incidence is biased. All the bias estimates obtained from the standard SCCS method without any age effect and the mis-specified SCCS method with 2 age groups are significantly different from zero. However, for the model with correctly specified age groups there is not enough evidence to reject the null hypothesis that there is no bias in estimating the exposure-related relative incidence value. Except when the true exposure-related relative incidence was 5 there was a borderline significant bias. For the correctly specified model, the absolute bias reduces with an increase in the number of cases used in the analysis and increases with an increase in the true relative incidence. The biases for the correctly specified SCCS are always smaller than the other two models. These results indicate that a new way of modelling the age effect that does not require age groups to be defined a priori is needed.

2.3.2 Semi-Parametric Model with Large Data Sets

Given that the parametric SCCS method can be sensitive to mis-specification of age groups, which may lead to biased estimates of the association between exposure and event outcome, the semi-parametric SCCS method, in which the age-specific relative incidence function is left unspecified, was proposed by Farrington and Whitaker (2006). However, the semi-parametric SCCS method faces computational problems with large data sets,

at least when fitted using standard software for log-linear models. We carried out a simulation study to investigate the computational demand of the semi-parametric method as the number of cases (and hence the number of parameters) in the model increases. We generated data using the same scenarios as in the previous section and fitted the semi-parametric model to each of the generated data sets. The number of cases simulated ranged from 10 to 478. Beyond this number of cases, the computer programs we used failed completely.

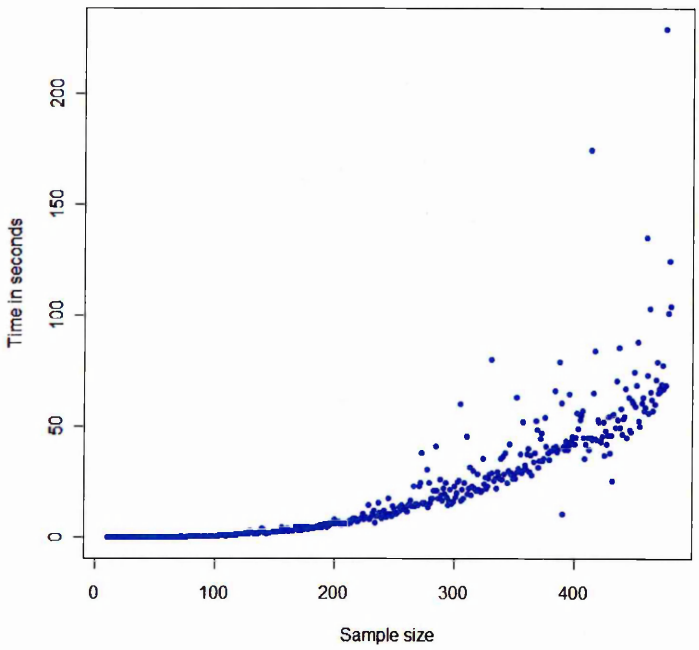

Figure 2.5: *Time elapsed to estimate parameters in the semi-parametric SCCS model against the sample size used.*

Figure 2.5 shows that the time elapsed to estimate parameters using the semi-parametric SCCS model increases as the number of cases in a study increases. It could not produce parameter estimates for data sets with greater than 478 cases. This is because for each interval $(a_i, b_i]$, $i = 1, 2, \ldots, N$ and each distinct event time, an indicator variable is defined

to record whether that time lies within the interval, so that the data set is expanded to $O(N^2)$. This typically causes capacity problems when N is above 478 or so. In addition, the number of age parameters to be estimated is of order N.

2.4 Discussion

The self-controlled case series method, developed to estimate the relative incidence of an acute event following exposure to vaccines, has been described in this chapter. The method uses information only from individuals with an adverse health event and implicitly controls all measured and unmeasured fixed confounding variables but time-varying covariates should be included in the model. The two versions of the method; the standard and the semi-parametric SCCS were introduced and their limitations investigated.

The simulation studies showed that the standard SCCS method may lead to biased exposure related relative incidence estimates and the semi-parametric method fails to fit for large data sets.

The SCCS method has witnessed considerable methodological development aimed at weakening the assumptions it requires. Thus, methods have been developed to handle event-dependent exposures and deaths (Kuhnert *et al.*, 2011), dependent recurrences (Farrington and Hocine, 2010), event dependent observation periods (Farrington *et al.*, 2011). The method has also been extended to the prospective monitoring of vaccine safety (Musonda *et al.*, 2008b; Hocine *et al.*, 2009). Escolano *et al.* (2013) extended the SCCS method for analyzing spontaneous reports of adverse events after vaccination aiming at rapid evaluation of a risk. When the timing of exposure onset is not known precisely, Mohammed *et al.* (2012) proposed to extend the SCCS method to take measurement errors into account. For this method, measurement error corrected SCCS, Mohammed *et al.*

(2013) developed a method that determines power and sample size. Simpson *et al.* (2013) extended the standard SCCS to include multiple time-varying confounding exposures (drugs) and their interactions, from a large-scale longitudinal observational database. Choice of optimal risk windows in SCCS vaccine safety studies has been proposed by Xu *et al.* (2013).

To address the limitation of the standard SCCS method the semi-parametric SCCS method was proposed. However no extension has been developed to avoid the limitation of the semi-parametric SCCS that it may suffer from computational problems as the number of cases in a study increases as shown in the simulation study. Moreover, the exposure effect in both the standard and the semi-parametric SCCS methods is represented by a step function. Step functions are known to provide a rather crude approximation of the true relationship. Therefore, the parametric way of modelling the exposure effect may be sensitive to mis-specification of exposure groups. Hence new ways of modelling age and exposure effects that are not sensitive to mis-specification and have no computational problems are required. In Chapter 4 we introduce smooth functions which are some of the possible ways of avoiding the limitations associated with the standard and semi-parametric SCCS methods.

Chapter 3

Review of SCCS Vaccine Studies

The SCCS method has been applied both in vaccine, non-vaccine pharmacoepidemiology and other areas of epidemiological studies. In this chapter we review how the SCCS method has been used in vaccine studies since its publication in 1995 and highlight good practice. We attempt to give some clear direction on how the method should be used and reported. Some misconceptions about the method and how it relates to other case-only study designs are clarified and some guidelines on reporting SCCS studies are given. However, our aims fall short of developing fully-fledged guidelines on reporting SCCS studies, which require detailed consideration of other applications in pharmacoepidemiology. Nevertheless, we hope that this review will contribute towards the eventual elaboration of such guidelines. This review has been published in Weldeselassie *et al.* (2011).

The chapter has three sections. In Section 3.1 our review criteria and methods are described. In Section 3.2 we present the results of our review, including specific discussion on: data description and accuracy, choice of observation and risk periods, potential biases, comparison of SCCS with other methods such as cohort and case-control, methodological issues, sensitivity analyses, software and good practice. Where appropriate, we also

include general comments about the method and make recommendations. Section 3.3 is a brief discussion of our findings and areas for further research on the SCCS method.

3.1 Methods

We identified SCCS studies which included a vaccine as an exposure, first published (in print or electronically) between 1995, when the SCCS method was first introduced, and the beginning of 2014. We identified papers by searching for those citing references Farrington (2004); Farrington *et al.* (1996); Farrington (1995); Farrington and Whitaker (2006); Whitaker *et al.* (2006, 2009); Andrews (2002); Musonda *et al.* (2006) in the following databases : Scopus, JSTOR, Science Direct, British Library and all those within the ISI Web of Knowledge.

Methodological papers were excluded, unless they included a specific application using SCCS not reported elsewhere, and sufficient detail of this application was provided.

Each paper was reviewed against a standard form which was piloted on 13 papers (see Appendix A). The form included details on: vaccines and adverse events studied, data collection and description, study population, sample size, observation period, age groups, the allowance for any other temporal confounders, risk periods and their rationale, sensitivity analyses undertaken, statistical features, reporting of results, whether key SCCS assumptions were met, any good, bad or unusual practice, and comparison with other study methods used in addition to SCCS.

3.2 Results

We identified 84 studies which met our selection criteria, four of them (Ali *et al.*, 2005; Burwen *et al.*, 2006; Farrington *et al.*, 1995; Gold *et al.*, 2010) were papers with a

methodological flavour, aimed at validating a surveillance system, but including a specific SCCS application. There were three notable exclusions. The first planned to use the SCCS method to study a possible association between vaccination and acute cerebellar ataxia (van der Maas *et al.*, 2009). However, that analysis was not undertaken owing to sparseness of the data, and for this reason was excluded. Two further papers (France *et al.*, 2004; Klein *et al.*, 2010) were excluded because, while referencing the SCCS literature, it was not clear that they intended to use it, and instead used a before and after vaccination design. As it turns out, this is in fact a special case of the SCCS design; we shall return to this issue later in the chapter. The papers were excluded because the authors could not be expected to report the study as if it were a SCCS study.

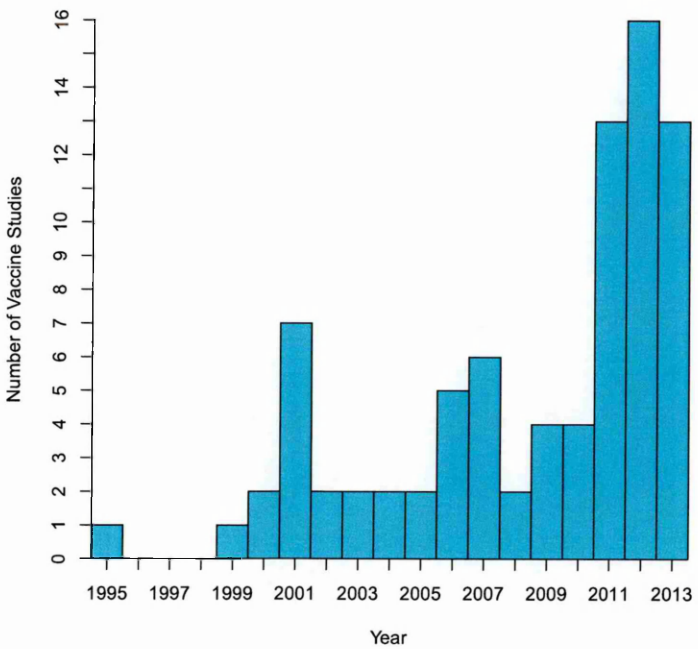

Figure 3.1: *Distribution of vaccine studies using self-controlled case-series by year of publication.*

Figure 3.1 presents the distribution by year of publication of these 84 studies (refer-

ences Andrews *et al.* (2001) and Sardinas *et al.* (2001) appeared in 2002, even though the journals are dated 2001; reference Gwini *et al.* (2011) was published electronically in 2010). Thirty-eight of the 84 papers appeared in 2000-2010 and 42 papers appeared in 2011–2013; Figure 3.1 suggests a moderate increase over the period 2000–2010 and a big increase in the period 2011–2013. There were also two studies until February in 2014.

Vaccines and Adverse Events Studied

Table 3.1 presents the vaccines and the adverse events studied between 1995 and 2010. For ease of presentation, adverse events have been grouped, as for example purpura (which includes idiopathic thrombocytopenic purpura (ITP), allergic purpura, and other purpura). Similarly, vaccine types (e.g. intranasal and parenteral influenza vaccines) have been listed under the same heading. Measles/mumps/rubella (MMR) and other measles-containing vaccines were the most frequently studied (17 studies), followed by influenza vaccines (13 studies) and vaccines containing pertussis antigens (eight studies). The sample sizes (cases or numbers of events) included in SCCS analyses ranged from the very small (only seven events in one analysis in Farrington *et al.* (1995)) to the very large (8,180 cases in Gwini *et al.* (2011), 22,400 in Smeeth *et al.* (2004)).

Typically, several vaccines and/or adverse events were studied at the same time. One study (Payne *et al.*, 2007) investigated concurrency of vaccination (administration of at least two vaccines on the same or adjacent days) as a risk factor. When several vaccines potentially related to the same outcome are administered at similar ages, their effects should be studied within the same model, as was done in Andrews *et al.* (2007). This also applies to non-vaccine exposures, as with influenza vaccination and influenza-like illness in GuillainBarre' syndrome (GBS) (Stowe *et al.*, 2008).

Table 3.1: *Vaccines and adverse events studied*

Vaccine	Adverse effect	Reference
Any concurrent	Hospitalization	Payne *et al.* (2007)
DT, Td	Convulsion	Andrews *et al.* (2007)
	Myocardial infraction, stroke	Smeeth *et al.* (2004)
DTP, DTaP	Convulsion	Andrews *et al.* (2007); Gold *et al.* (2010); Farrington *et al.* (1995); Huang *et al.* (2010)
	Encephalitis	Ward *et al.* (2007)
	Immune haemolytic anaemia	Naleway *et al.* (2009)
	Wheeze onset	Mullooly *et al.* (2002)
DTP/Hib/IPV	Apnoea, convulsion, crying, diarrhoea, feeding problem,fever	Andrews *et al.* (2010)
HBV	Demyelination onset	Hocine *et al.* (2007)
	Immune haemolytic anaemia	Naleway *et al.* (2009)
	Wheeze onset	Mullooly *et al.* (2002)
Hib	Wheeze onset	Mullooly *et al.* (2002)
Influenza	Asthma exacerbation	Kramarz *et al.* (2000, 2001); Tata *et al.* (2003)
	Bells palsy	Mutsch *et al.* (2004); Stowe *et al.* (2006);
	Cellulitis or abscess, UTI	Burwen *et al.* (2006)
	COPD exacerbation	Tata *et al.* (2003)
	Gastritis/duodenitis	Hambidge *et al.* (2006)
	GuillainBarre' syndrome	Hughes *et al.* (2006); Stowe *et al.* (2008); Juurlink *et al.* (2006);

Table 3.1: Continued

Influenza	Myasthenia gravis	Zinman *et al.* (2009)
	Myocardial infraction, stroke	Gwini *et al.* (2011); Smeeth *et al.* (2004)
MCCV	Convulsions, purpura	Andrews *et al.* (2007)
	Encephalitis	Ward *et al.* (2007)
	Nephritic syndrome relapse	Taylor *et al.* (2007)
Measles	Acute respiratory tract infection, arthropod-borne, viral fever, gastroenteritis, pneumonia, tonsillitis	Ali *et al.* (2005)
	Autism	Farrington *et al.* (2001); Taylor *et al.* (1999);
MMR	Aseptic meningitis	Dourado *et al.* (2000); Miller *et al.* (2007); Farrington *et al.* (1995);
	Autism	Andrews *et al.* (2002); Taylor *et al.* (1999); Farrington *et al.* (2001);
	Bacterial or viral infection	Miller *et al.* (2003); Stowe *et al.* (2009);
	Convulsion	Andrews *et al.* (2007); Miller *et al.* (2007); Farrington *et al.* (1995); Gold *et al.* (2010);
	Encephalitis	Ward *et al.* (2007)
	Gait disturbance	Miller *et al.* (2005)
	Purpura (including ITP)	Andrews *et al.* (2007); Farrington *et al.* (1995); France *et al.* (2008); Gold *et al.* (2010); Miller *et al.* (2001); Stowe *et al.* (2001);
	Wheeze onset	Mullooly *et al.* (2002)

Table 3.1: Continued

OPV	Intussusception	Andrews *et al.* (2001); Cameron *et al.* (2006); Sardinas *et al.* (2001)
	Wheeze onset	Mullooly *et al.* (2002)
Pneumococcal	Bells palsy	Stowe *et al.* (2006)
	Cellulitis or abscess, UTI	Burwen *et al.* (2006)
	GuillainBarre' syndrome	Stowe *et al.* (2008)
	Myocardial infarction, stroke	Smeeth *et al.* (2004)
Rotavirus	Intussusception	Murphy *et al.* (2001)

COPD, Chronic obstructive pulmonary disease; DT, diphtheria/tetanus vaccine; DTaP, diphtheria/tetanus/acellular pertussis vaccine; DTP, diphtheria/tetanus/pertussis vaccine; HBV hepatitis B virus vaccine; Hib, Haemophilus influenzae type b vaccine; IPV, inactivated poliovirus vaccine; ITP, idiopathic thrombocytopenic purpura; MCCV, meningococcal group C conjugate vaccine; MMR, measles/mumps/rubella vaccine; OPV, oral poliovirus vaccine; Td, tetanus/diphtheria booster vaccine; UTI, urinary tract infection.

In 22 studies, vaccines were given in multiple doses; in 12 of these, dose-specific effects were investigated. The SCCS method can only cope with a single outcome variable at a time. The most frequently studied events were convulsion (including febrile convulsion and aseptic meningitis) and purpura (six studies each). There were four studies of intussusception, and three each of autism and GBS.

Data Description and Data Accuracy

All the 84 studies were felt to provide sufficient detail of how the data were collected, so that it would be possible to see any dependence between ascertainment of cases and vaccination history. Out of the 40 studies in period 1995–2010, 16 obtained data on vaccinations and outcomes from a single database [of these, seven studies used the United Kingdom's General Practice Research Database (GPRD) and six the United States' Vaccine Safety Datalink], 14 linked two or more databases, and in 10 data were obtained from other sources.

Case-note reviews were undertaken in 18 studies, and in two of these the review was commendably reported as blinded to vaccine history. In one study (Andrews *et al.*, 2007), case notes were used to identify vaccinations. This may bias results towards a positive association, in as much as vaccinations prior to the event are more likely to be ascertained by case-note review than vaccinations after the event. However, in this study the association was not significant, and so the ascertainment procedure in this instance leant further weight to the conclusions reached. Most studies had full information on the day of vaccination and the day of event (studies Farrington *et al.* (2001) and Taylor *et al.* (1999) used month as the time unit for analysis, but with long risk periods). In Burwen *et al.* (2006) and Dourado *et al.* (2000), dates of vaccination were imputed rather than observed exactly. The sensitivity to imputation errors depends on the lengths of the risk periods used, and it would be advisable to study this by sensitivity analyses, although none were reported. In Andrews *et al.* (2010), vaccination dates were known exactly, but the types of vaccines used at different times were derived indirectly.

Observation Periods and Risk Periods

A well-conducted SCCS study requires great rigour in the definition of observation periods and risk periods for each case. In all 84 studies, observation periods were defined with sufficient detail to reconstruct the study. The idiosyncrasies of specific databases need to be allowed for appropriately in defining observation periods. Thus, some studies excluded day of vaccination (or allocated it a special parameter) owing to the fact that, in some information systems, past events are retrospectively recorded on day of vaccination; left uncorrected, this would induce spurious associations on the day of vaccination. This effect is illustrated graphically in marked fashion in Tata *et al.* (2003). In one study in the GPRD (Andrews *et al.*, 2010), events on the day of vaccination were validated by case-note review.

The choice of risk periods should be made a priori and its rationale explained. The risk periods were explicitly defined in all 84 studies. Typically, the choice of the risk periods is based on reference to previous studies or hypotheses, as in France *et al.* (2008) for example; on biologically plausible mechanisms (Farrington *et al.*, 1995) ; or by expert opinion (Miller *et al.*, 2005). Different risk periods may sometimes reflect different scientific questions. For example, in Farrington *et al.* (1995), the 6- to 11-day risk period post-MMR was chosen to capture febrile convulsions associated with the measles component of the vaccine, while the 15- to 35-day risk period was chosen to capture convulsions associated with the mumps component. Inevitably, in some circumstances the risk period is not known, and so the choice is arbitrary; if so this should be stated (Farrington *et al.*, 2001). Three studies: Farrington *et al.* (2001); Hocine *et al.* (2007); Kramarz *et al.* (2001) used indefinite post vaccination risk periods. In several studies (e.g. Juurlink *et al.* (2006)) a sensitivity

analysis was undertaken by varying the risk period. A further approach is to use several adjacent risk periods in the same analysis. For example, to investigate seizures and acellular pertussis vaccines, study Huang *et al.* (2010) used the risk periods 0 days (i.e. the day of vaccination) and 1-3 days after vaccination. When results are similar across risk periods, or when data are lacking, contiguous risk periods can be combined. When a relatively long risk period is used, it is advisable to undertake secondary analyses to identify clustering or otherwise of cases within that risk period. Examples include Miller *et al.* (2001), where a clustering of ITP cases was found 15–28 days post vaccination within the 42-day risk period studied, and Farrington *et al.* (2001), where no clustering of autism cases was found in adjacent 2-year intervals within the unlimited post-MMR risk period studied.

Confounders

SCCS studies adjust automatically for time-invariant multiplicative confounders. However, effect modification by fixed covariates can be investigated through interactions with the vaccine effect : for example in Gwini *et al.* (2011) such effects were investigated, for sex and age at start of observation. The SCCS method, in common with other epidemiological methods, is prone to bias from uncontrolled age- or time-varying confounders. In vaccine studies, particularly those undertaken in children, age (or in some cases season, or both) is likely to be the major confounder, and should, as a rule, be adjusted for in the analysis, unless observation periods are extremely short. Seven studies from 1995–2010 did not report using any kind of temporal adjustment; in four of these, the observation period was less than a year. Of the remaining 33 studies, 19 adjusted for age only, three for season only, one for calendar time only, six for age and season, one for age and calendar

time, and two for age, season and calendar time (e.g. Huang *et al.* (2010)).

Only one study (Hocine *et al.*, 2007) used the semi-parametric model (Farrington and Whitaker, 2006), in which it is not necessary to specify age classes. If a standard method of age adjustment is used, it is good practice to check that the age model used is adequate, by varying the number of age classes used. Two studies reported such sensitivity analyses (Hocine *et al.*, 2007; Smeeth *et al.*, 2004). One study (Hughes *et al.*, 2006) controlled the time-varying confounding variable age as a continuous covariate and one study Mullooly *et al.* (2011) used linear and quadratic functions, although no details of how these were achieved were given; such a method of control is not straightforward owing to the conditioning (see Chapters 5 and 7 on how age can be controlled as a continuous variable). Carlin *et al.* (2013), used fractional polynomials to control for age in addition to the standard method.

Control for age-varying or time-varying confounders other than age or season require the confounder to be measured over time. For example, in an analysis of influenza vaccine and GBS (Stowe *et al.*, 2008), the authors controlled for the possible confounding effect of influenza-like illness. However, it is often impractical to measure time-varying confounders. For example, the healthy vaccine effect is a form of confounding by an unmeasured time-varying factor. This affects SCCS studies as well as other study designs. The potential impact of such bias therefore requires careful discussion.

Discussion of Potential Biases

The three key assumptions of the SCCS method listed in Section 2.1.1 of Chapter 2 should be checked, as far as possible, and discussed. We consider these three assumptions in turn.

Assumption (1) : that the events are either recurrent and independent within individuals, or non-recurrent and uncommon, is not usually problematic. For recurrent events, sensitivity to the independence assumption can readily be tested by restricting the analysis to first events, provided these are uncommon in the population considered; see Farrington and Whitaker (2006); Whitaker *et al.* (2006) for an example with MMR and ITP. More complex approaches to correcting for non-independence of recurrent events are discussed in Farrington and Hocine (2010). Simulation studies in Farrington *et al.* (2011) show that the bias is negligible when the risk that an unvaccinated individual will experience an event over the observation period is under 10%. Most adverse events of interest in post-licensure studies are much less common than this.

Assumption (2) : that the event should not affect the subsequent probability of vaccination, is perhaps the most important for vaccine studies. This assumption fails if the event is a contra-indication for vaccination (as with intussusception and rotavirus vaccination since the publication of Murphy *et al.* (2001)), or if vaccination after the event is more or less likely (as with GBS and influenza vaccination). A third possibility is that vaccination is deferred after (or more rarely, precipitated by) an event, so that the impact of the event on vaccination is short-lived. Nevertheless, an important feature of such biases is that their direction is predictable: if the event reduces the probability of subsequent vaccination, then the relative incidence associated with vaccination will be biased upwards. This is because vaccinations are then less likely to arise after the event. There are three main ways of coping with such bias: including pre-vaccination risk periods to allow for short-term deferral of vaccination (or indeed to investigate the presence of longer term effects); exclusion of all pre-vaccination time (so that the observation period begins with vaccination), which works provided the vaccine can only be given at most

once during the projected observation period; and the use of more complex analytic techniques (Farrington *et al.*, 2009). Of the studies reviewed, 16 used pre-vaccination risk periods (see, e.g. Andrews *et al.* (2001) and Sardinas *et al.* (2001)), and three (Juurlink *et al.*, 2006; Smeeth *et al.*, 2004; Zinman *et al.*, 2009) started observation at vaccination for some analyses. An extended version of the SCCS method (Farrington *et al.*, 2009), that allows censored, perturbed or curtailed post-event exposures was applied to investigate the association between Guillain Barre' Syndrome and influenza vaccines by three studies Dodd *et al.* (2013); Galeotti *et al.* (2013) and Romio *et al.* (2014). Traversa *et al.* (2011) used the extended method to investigate if sudden unexpected deaths were associated to vaccinations during the first two years of life.

Assumption (3) : that the observation periods are not event-dependent, may be violated, for example, if events increase short-term mortality, or the event of interest is death. This was not an issue in any of the 84 studies reviewed. SCCS methods for dealing with such situations are discussed in Farrington *et al.* (2011, 2009); Kuhnert *et al.* (2011).

Comparisons With Other Statistical Methods

In addition to implementing the SCCS method, more than 12 studies used or reported results obtained on the same data using other study designs. These included cohort, case-control, and ecological methods. The different methods should produce the same results, provided that all confounding has been controlled and that the assumptions required are met.

Using several methods of analysis is recommended, as it can reinforce conclusions or shed light on possible sources of bias, when these differ for different study designs. Table 3.2 presents the results obtained using SCCS and other methods, for a selection of

analyses.

Table 3.2: *Selected relative incidence (RI) estimates from self-controlled case series method and RI or odds ratio (OR) from other study designs applied to the same case data, and 95% confidence interval (CI)*

Vaccine (adverse effect) (reference)	SCCS RI (95% CI)	Other study type	RI or OR (95% CI)
MMR (aseptic meningitis) (Dourado *et al.*, 2000)	30.4(11.5–80.8)	Before/after ecological analysis	14.3(7.9-25.7
MMR (ITP) (France *et al.*, 2008)	5.38(2.72–10.62)	Cohort	3.94(2.01-7.69)
Influenza (gastritis/duodenitis) (Hambidge *et al.*, 2006)	4.54(1.90-10.86)	Case crossover*	4.33(1.23-15.21)
HBV (first demyelination) (Hocine *et al.*, 2007)	1.68(0.77-3.68)	Case-control	1.8(0.7-4.6)
DTaP (seizure) (Huang *et al.*, 2010)	0.91(0.75-1.10)	Cohort	0.87(0.72-1.05)
Influenza (asthma exacerbation) (Kramarz *et al.*, 2000)	0.98(0.76-1.27)	Cohort	1.39(1.08-1.77)
Influenza (asthma exacerbation) (Kramarz *et al.*, 2001)	0.65(0.52–0.80)	Cohort	1.4(1.2-1.5)
HBV (wheezing onset) (Mullooly *et al.*, 2002)	0.41(0.24–0.70)	Case-control	0.59(0.22–1.59)
Oral rotavirus (intussusception) (Murphy *et al.*, 2001)	29.4(16.1–53.6)	Case-control	21.7(9.6-48.9)
Intranasal flu vaccine (Bells palsy) (Mutsch *et al.*, 2004)	35.6(14.1–89.8)	Case-control	84.0(20.1-351.9)
Concurrent vaccines (hospitalization) (Payne *et al.*, 2007)	Identical	Cox regression	0.90(0.75-1.09)
MCCV (nephritic syndrome relapse) (Taylor *et al.*, 2007)	0.95(0.61-1.47)	Before/after ecological analysis	1.05(0.95-1.15)

For abbreviations refer Table 3.1 note.

* This description is incorrect : it is actually another SCCS (see text).

The results obtained using SCCS were broadly similar to those obtained by other

methods, with the exception of studies of influenza vaccine and asthma exacerbation (Kramarz *et al.*, 2000, 2001) where the SCCS method found a protective or null effect, but a cohort analysis found a positive association. The most likely explanation for this discrepancy is residual indication bias in the cohort study, children with more severe asthma being more likely to receive influenza vaccine. In the cohort study, underlying asthma severity was quantified using available proxy variables; self-control in the SCCS study was arguably more effective in correcting for indication bias. More generally, the results of a SCCS study should be unaffected by unmeasured or incompletely controlled confounders, and in this sense ought to be more reliable, provided that the assumptions of the method are satisfied.

In a study of hepatitis B vaccine (HBV) and wheezing onset (Mullooly *et al.*, 2002), the point estimates from SCCS and a case-control study were of the same order, but the greater precision of the SCCS method in this case produced a statistically significant effect. The better precision of the SCCS method was also noted in another study of HBV (Hocine *et al.*, 2007), where it was pointed out that some cases cannot be used in matched case-control studies owing to lack of matching controls ; the SCCS method does not suffer from this problem. In one study (Hambidge *et al.*, 2006) the alternative method was incorrectly described as a case-crossover design, when in fact it was another SCCS with a before and after vaccination observation period. The distinction between SCCS and case-crossover methods (Delaney and Suissa, 2009) stems from the fact that, as described above, SCCS studies are based on cohort designs, whereas case-crossover studies are based on case-control designs. The use of case-crossover methods for vaccine safety studies is discussed briefly in Farrington (2004).

The SCCS method is never exactly as powerful (and therefore, does not yield as precise

estimates) as a cohort study with the same cases, unless, as often occurs in practice, there is unexplained between individual variation in the cohort study which inflates the uncertainty. However, when risk periods are short relative to observation periods, the power of the SCCS method approaches that of a cohort study. However, SCCS studies with long or indefinite risk periods (Farrington *et al.*, 2001) may have substantially lower power than a cohort study with the same cases (see the discussion of Farrington and Whitaker (2006)). A SCCS study is usually more powerful than a case-control study with the same cases and with a single control per case (Farrington *et al.*, 1996). (As the number of controls increases, the power of the case-control study increases.)

Methodological Issues

An unusual feature of the SCCS method is that post-event time is included in the analysis. This is a consequence of the fact that the method works by conditioning, for each individual, on that person's vaccination history over the entire observation period, and on the number of events arising within that period. It follows that observation time should not be censored at the event. One study (Hughes *et al.*, 2006) did censor observation at the event, in this instance GBS, ostensibly because patients who have had GBS may be advised not to have further immunizations. If GBS patients are less likely to receive immunizations after experiencing the adverse event then, as noted above, a standard SCCS analysis would have resulted in an overestimate of the relative incidence. Censoring at event, however, produces bias of unpredictable direction, and is not recommended.

Several studies of potentially recurrent events, such as convulsions (Huang *et al.*, 2010), ITP (Miller *et al.*, 2001) or GBS (Stowe *et al.*, 2008), considered repeat events to be part of the same episode if separated by less than some minimum time period τ. This presents

the methodological problem that, after an event, no other event can then occur for a time interval τ: an instance of immortal time, which, if included in the analysis, may result in bias (Suissa, 2007). Generally τ is short and repeat events are relatively uncommon, so any such bias is likely to be small. A simple approach is to perform a sensitivity analysis restricted to first events, which also sidesteps the requirement for repeat episodes to be independent. One interesting study (France *et al.*, 2008) excluded person-time for a period τ after each episode; however, the performance of such a strategy requires further investigation.

Several SCCS studies defined observation periods relative to the day of vaccination, either starting with vaccination and ending a fixed number of days after vaccination (Juurlink *et al.*, 2006; Zinman *et al.*, 2009), or starting and ending some fixed number of days before and after vaccination (Ali *et al.*, 2005; Burwen *et al.*, 2006); we refer to such studies as before and after designs. For some studies this was done for convenience of data collection. While not invalid, this approach results in short observation periods, which is not optimal, as information from events occurring at other times is not used. In addition, the short control periods may only include time when the risk of temporal bias is high. For example, bias from delayed vaccination following an event may artificially depress the incidence in the period immediately preceding vaccination. This effect is very apparent on the plots of intervals between vaccination and events in Burwen *et al.* (2006), which shows a marked trough of hospitalizations in the week preceding vaccination (this week was, rightly, excluded from the analysis).

As explained in Section 2.1.1 of Chapter 2, the SCCS method is derived from a cohort model by conditioning on the number of events observed, as well as on vaccination history. Thus, a conditional (Poisson) model is used to estimate the parameters. Fewer than half

of the 40 studies in 1995–2010 indicated that a conditional Poisson regression model was used, either explicitly (e.g. (France *et al.*, 2008; Gwini *et al.*, 2011; Hambidge *et al.*, 2006)) or with words to that effect (as in Gold *et al.* (2010); Zinman *et al.* (2009)). In a few studies it was unclear whether a conditional or unconditional model was fitted (e.g. (Payne *et al.*, 2007)). The only circumstance in which an unconditional Poisson model (i.e. one in which the number of events per individual is not regarded as fixed) may be used in a SCCS analysis is when all individuals have identical observation periods and vaccination histories. In this special case, the conditional and unconditional methods give the same results. In two further instances, the method of analysis appeared somewhat idiosyncratic (Burwen *et al.*, 2006; Hughes *et al.*, 2006).

Useful Plots

Several studies (e.g. Dourado *et al.* (2000); Stowe *et al.* (2009); Tata *et al.* (2003)) included plots showing the intervals between events and vaccination; these are useful for visualizing the association between exposure and event (although they are also prone to censoring effects), and for identifying pre-vaccination troughs. Such plots are trickier to draw for multi-dose vaccines, but are useful nonetheless (Murphy *et al.*, 2001). Other studies (e.g. France *et al.* (2008); Naleway *et al.* (2009)) illustrated the case ascertainment procedure using a flow diagram, which presents clearly the inclusions and exclusions applied to assemble the cases, and hence can help the reader assess any biases that may have arisen in the process. Further useful plots include those illustrating the risk periods used (Gwini *et al.*, 2011; Smeeth *et al.*, 2004), those showing estimated age or season effects (Hocine *et al.*, 2007) and, for complex analyses with many endpoints, graphical representation of the relative incidences (Andrews *et al.*, 2010).

Power and Sample Size Issues

In studies involving very uncommon events, power and sample size considerations are particularly important (Musonda *et al.*, 2006). One study (Gwini *et al.*, 2011) reported checking the sample size required to achieve 90% power to detect at least a doubling of risk. The relevant sample size is the number of events, and if this is too small the estimates and confidence intervals may not be accurate. To aid interpretation, it is important to report the numbers of events in risk and control periods. The larger the imbalance in the expected numbers of events in the risk and control periods, the worse the small sample bias. This is most likely to affect studies with very short risk periods. Simulation studies reported in Musonda *et al.* (2008a) suggest that the small sample bias is likely to be small provided at least 2.5 events are expected in the risk period. Note also that a small sample size may adversely affect the ability to control effectively for the effect of age and other time-varying confounders.

Sensitivity Analyses

Sensitivity analyses have been mentioned throughout this chapter. They provide a simple way of evaluating the robustness of the results; we focus here on where they may be used (other useful sensitivity analyses than those described here can doubtless be performed).

When the SCCS model is used with parametric adjustment for age we recommend checking the sensitivity of exposure risk estimates to choice of age group, by increasing the number of age groups (Hocine *et al.*, 2007; Smeeth *et al.*, 2004).

Sensitivity analyses of risk periods should be motivated explicitly (as in Juurlink *et al.* (2006)). Researchers may also wish to consider whether it would be sensible to explore

sensitivity of results by adding washout periods to the chosen risk period, removing the day of vaccination or including pre-vaccination risk periods.

If recurrent adverse events occur in episodes, and there is a lack of clarity over whether repeat events are part of the same episode, sensitivity to the choice of definition of episodes can be checked. Note that analyses of first events only can be carried out to avoid any issue of lack of independence between adverse events.

If exact dates or timings of exposures or events are unknown and have to be imputed, sensitivity to how these timings are imputed should be explored.

When sensitivity analyses are performed, they should be reported, with full details when they relate to risk periods, washout periods and pre-vaccination periods. It is important to distinguish between them and the pre-planned primary analyses. If sensitivity analyses suggest possible departures from the assumptions of the method, this should be stated explicitly. If it is thought that departure from assumptions might affect the results, then, where possible, alternative methods of analysis should be used in conjunction with SCCS.

Software for SCCS Analyses

Twelve of the studies that appeared in 1995–2010 reported which statistical package was used to undertake the SCCS analysis. Six used Stata (StataCorp, USA), five used SAS (SAS Institute Inc.,USA) and one used GLIM (NAG, UK). Further information about fitting SCCS models using these packages and other standard softwares may be found in Whitaker *et al.* (2006) and on the associated website (http://statistics.open.ac.uk/sccs).

The SCCS model is most conveniently fitted using software designed for Poisson regression models with fixed effects (in this case, the levels of the fixed effects represent

distinct individuals). This method of fitting the models exploits a convenient technical fact known as the 'Poisson trick', whereby a multinomial likelihood (which applies for the SCCS method, see Farrington (1995); Farrington and Whitaker (2006)) can be maximised using a Poisson model. However, this trick has its limits : for example, fitting age as a continuous variable cannot be done in this way, because it does not allow for the fact that age varies within each risk or control interval, see Chapters 5 and 7 on how to control for age as a continuous variable.

3.3 Discussion

Review of vaccine studies that made use of the SCCS method from 1995 to the beginning of 2014 was done in this chapter. The review was based on papers quoting key papers on the case-series method. We are aware of several independent reinventions of the SCCS method in different contexts : the bidirectional case-crossover method applied to fixed observation times (Navidi, 1998) and the time-stratified case-crossover approach (Lumley and Levy, 2000), developed for the analysis of environmental time-series data (see Vines and Farrington (2001); Whitaker *et al.* (2007), for a discussion of the connections with the SCCS method), and the method of Becker *et al.* (2004) applied to venous thromboembolism after long-haul flights. None of these versions of SCCS have so far been used in connection with vaccine safety. Thus, to the best of our knowledge, we have included all applications of SCCS methodology to vaccine studies that appeared by the beginning of 2014.

We identified and reviewed 84 papers which applied the SCCS method to vaccine studies. In general the method was applied appropriately. All 84 studies provided sufficient detail of how their data were collected, which enabled the reader to make sure that events

are identified independently of vaccinations. Moreover, observation and risk periods were generally carefully specified. Most studies adjusted for age and/or season as appropriate.

The following key issues emerge when using the SCCS method. Ascertainment of cases and collection of data on exposure history should be independent, as bias may result if case ascertainment was influenced by knowledge of exposure status. The observation and risk periods should be clearly defined, and the choice of risk period should be justified. Where necessary, age and season effects should be allowed for, and when using the standard model, sensitivity to the choice of age and seasonal groups should be checked. Other relevant time-varying covariates (such as concurrent vaccinations and other exposures) which may be associated with both the exposure and outcome should be identified and, if possible, taken into account in the analysis. The validity of the assumptions required by the SCCS method should be carefully considered and appropriate supplementary sensitivity analyses undertaken where these come into question.

A few papers suggest there remains a degree of confusion about what a SCCS study entails, in particular how it differs from a before and after vaccination analysis or from the case-crossover paradigm. This is wholly unsurprising, owing to the somewhat abstruse and technical, yet fundamental, distinction between conditional and unconditional analyses. In recent methodological paper Glanz *et al.* (2006) a before and after design is described, described as a risk interval method, which is in fact a special case of a SCCS design. The term case centred has also been used to describe such designs (Klein *et al.*, 2010). We excluded two papers France *et al.* (2004); Klein *et al.* (2010) with before and after analyses from our review because they did not describe the design as SCCS; several before and after analyses that did were included in the review. In fact, all these studies are special cases of the SCCS design. Nevertheless, the picture that emerges is dominated by

the numerous impressive and often imaginative applications of the method.

This review has raised some further methodological issues worthy of further study. One such is how best to handle the immortal time after an event, during which recurrences are classified as part of the same episode, and whether ignoring this effect has any substantive bearing on the results. Another is to study and quantify the bias that results from censoring observation periods at events. Sensitivity analyses may be indicated in both circumstances. Further, while the SCCS method is only applicable with a single outcome variable at a time, it may be desirable to study several outcomes jointly. A bivariate SCCS method has been suggested for the analysis of antibiotic resistance (Hocine *et al.*, 2009); perhaps similar ideas can be used for a multivariate SCCS applied to vaccine safety, in which several possibly dependent outcomes could be studied at the same time.

SCCS is a relatively new statistical methodology, and the issues that require particular emphasis and care in reporting have, therefore, only become apparent over time. The development of suitable guidelines for reporting such studies, in vaccine safety and pharmacoepidemiology more widely, may perhaps now be indicated.

All but two studies applied the standard SCCS method where the age and exposure effects are represented by piecewise constant step functions. However, the standard SCCS method has a limitation that misspecification of age groups might result biased estimates as presented in section 2.3 of Chapter 2 and only few of the reviewed studies did sensitivity analysis. Therefore, it worth modelling the age and exposure related relative incidence functions by smooth functions that avoid the limitations of the standard SCCS methods. In the next chapter we present some of the possible smooth function which could be used in the SCCS context.

Chapter 4

Basic Concepts of Smooth Functions

The standard self-controlled case series method, as described in Chapter 2, uses step functions to model the effects of age and exposure. Alternative ways of modelling, that avoid the use of step functions in general and their limitations in the SCCS method in particular, are presented in this chapter. In Section 4.1 we introduce polynomial functions followed by fractional polynomials in Section 4.2. Then we describe spline functions based on truncated power functions, B-splines, M-splines and I-splines in Section 4.3 followed by a discussion in Section 4.4

4.1 Polynomial Functions

The simplest way of replacing a step function with a smooth function is to use a polynomial of degree higher than zero. Polynomial function $f(t)$ can be constructed as a linear combination of functions, $f(t) = \sum \alpha_l h_l(t)$, where $h_l(t)$ are known as basis functions.

For example the basis functions of a straight line model, $f(t)$, are $h_0(t) = 1$ and $h_1(t) = t$, where t is the variable of interest. Then $f(t)$ is expressed as a weighted sum of

the basis functions, i.e

$$f(t) = \sum_{l=0}^{1} \alpha_l h_l(t) = \alpha_0 + \alpha_1 t.$$

For this linear function the design matrix is:

$$T = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}$$

where $t_1, t_2, \ldots, t_n$ are the observations on t and n is the number of observations. These basis functions are graphically shown on the top row in panel a of Figure 4.1. On the top panels of the figure, the basis functions are denoted by solid lines and the corresponding curves fitted using these basis functions are presented on the bottom panels as dashed lines. The 100 data points denoted by circles were simulated from a normal distribution with a mean of $sin(t)+2$ and a standard deviation of 0.5 and the solid line in the bottom panel shows the true curve. The domain of t ranges from 0 to 8.

To increase the flexibility of a polynomial function, that is to achieve a more flexible approximation of $f(t)$, we can increase the order of the polynomial function by adding more basis functions. For example, for a quadratic model, the basis functions are $h_0(t) = 1$, $h_1(t) = t$, and $h_2(t) = t^2$ and in general for a degree p polynomial model the basis functions are

$$h_0(t) = 1, h_1(t) = t, h_2(t) = t^2, h_3(t) = t^3, \ldots, h_p(t) = t^p$$

with a design matrix of

$$T = \begin{bmatrix} 1 & t_1 & t_1^2 & \dots & t_1^p \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & t_n & t_n^2 & \dots & t_n^p \end{bmatrix}.$$

Then

$$f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 \cdots + \alpha_p t^p,$$

where the coefficients $\alpha_0, \alpha_1, \alpha_2 \ldots, \alpha_p$ determine the shape of the function $f(t)$.

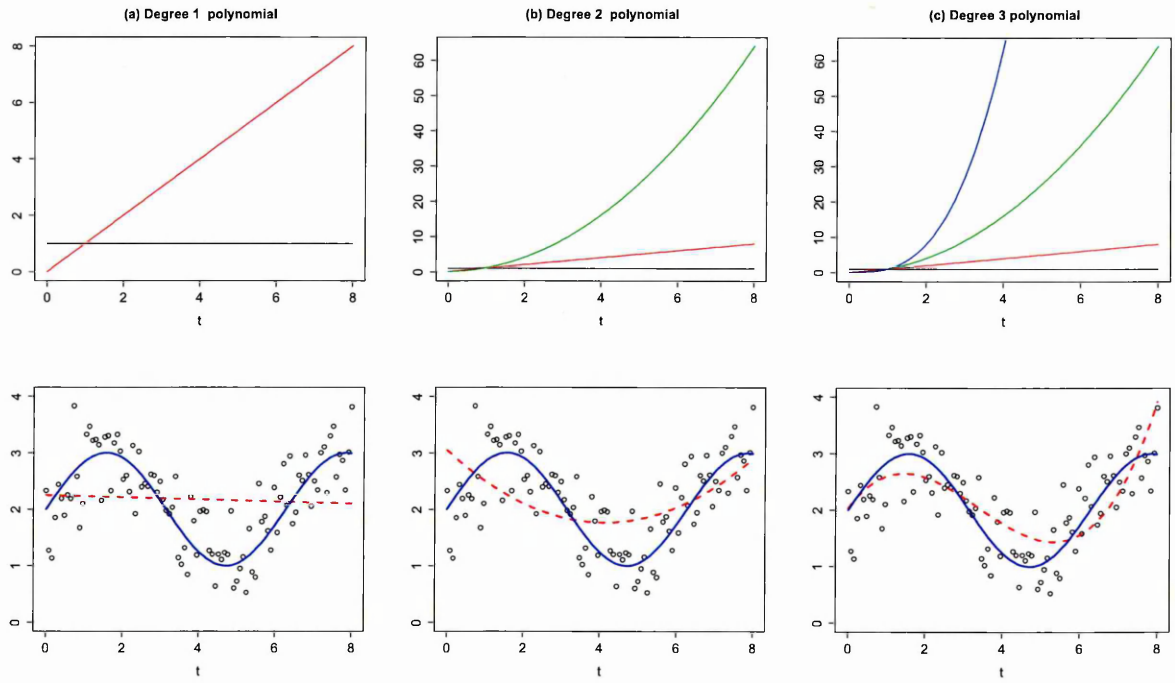

Figure 4.1: *Basis functions and fitted regression curves to 100 data points simulated from a normal distribution with mean* $sin(t) + 2$ *and standard deviation* 0.5. *Panels (a) represent a degree 1 polynomial, panels (b) a quadratic and panels (c) a cubic polynomial. The top row shows basis functions and the bottom row shows the fitted and true polynomial functions. In all the panels, data points are represented by circles, the true function is denoted by a solid line, the dashed lines denote the fitted polynomial curves*

Basis functions for quadratic and cubic regression curves are presented in Figure 4.1

on the top row of panels b and c respectively and their corresponding fitted curves are shown on the bottom row.

According to Ramsay and Silverman (1997), a good basis function should be chosen such that estimation of the coefficients α_l and computation of the basis functions is fast, they are flexible enough to exhibit the required curvature where needed and of course, nearly linear when appropriate. They should be differentiable as required i.e one or more of the derivatives of the approximation made based on the basis functions should behave reasonably. And the other property taken into consideration to make a choice among the basis functions is that they allow one to do constrained modelling when required, for example by satisfying monotonicity, positivity conditions etc.

4.2 Fractional Polynomials

Polynomial regression, although a popular technique, has limitations in the fact that individual observations can exert an influence in unexpected ways on remote parts of the curve (Green and Silverman, 1994). Low order models such as quadratics lack flexibility, and higher order fitted curves have a propensity to produce artefacts, such as waviness and end-effects (Royston and Altman, 1997).

To address these limitations a family of fractional polynomials was introduced by Royston and Altman (1994) as generalizations of the conventional polynomial class of functions by considering not only positive integer powers, but also negative and fractional powers. A fractional polynomial of degree m is defined as

$$f(t) = \sum_{l=0}^{m} \alpha_l H_l(t),$$

where m is a positive integer, and $H_l(t)$ is given by

$$H_l(t) = \begin{cases} t^{p_l}, & \text{if } p_l \neq p_{l-1} \\ H_{l-1}(t) \times \log(t), & \text{if } p_l = p_{l-1} \end{cases}$$

with $p_1 \leq p_2 \leq \cdots \leq p_m$ a sequence of powers and $p_0 = 0$. $H_0 = 1$, if $p_l = 0$ then $H_l(t) = t^{p_l}$ is taken, by definition, to be $H_l(t) = \log t$. Allowing non integer powers in fractional polynomials increases their flexibility even in lower order models and they fit better at extreme values of the observed range of covariates unlike conventional polynomials. Royston and Altman (1994) argued that, in practice, fractional polynomials of degree higher than 2 are rarely needed and suggested choosing the values of powers from a set of numbers between -2 and 3, i.e $\{-2, -1, -0.5, 0, 0.5, 1, 2, ..., \max(3, m)\}$. For fractional polynomial models with $m > 1$, e.g $m = 2$ if $p_2 = p_1$ the models do not degenerate to models with fewer powers (Royston and Altman, 1997). That is $\alpha_0 + \alpha_1 H_1(t) + \alpha_2 H_2(t)$ are not of the same degree as $\alpha_0 + \alpha_1 H_1(t)$ if $p_2 = p_1$. The model has three parameters (degree 2)

$$\alpha_0 + \alpha_1 t^{p1} + \alpha_2 t^{p_1} \log t.$$

To choose the best powers p_1 and p_2, when $m = 2$, fractional polynomial models with each possible pair of p_1 and p_2 are fitted and deviance values are computed. A model that gives the smallest deviance is chosen as the best model.

The basis functions for a function, $f(t)$, approximated by a fractional polynomial model with $m = 2$ are:

$$h_0(t) = 1, h_1(t) = t^{p_1}, h_2(t) = t^{p_2} \qquad \text{if } 0 \neq p_1 < p_2 \neq 0$$

$$h_0(t) = 1, h_1(t) = t^{p_1}, h_2(t) = t^{p_1} \log t \qquad \text{if } 0 \neq p_1 = p_2 \neq 0$$

$$h_0(t) = 1, h_1(t) = \log t, h_2(t) = t^{p_2} \qquad \text{if } p_1 = 0$$

$$h_0(t) = 1, h_1(t) = t^{p_1}, h_2(t) = \log t \qquad \text{if } p_2 = 0$$

$$h_0(t) = 1, h_1(t) = \log t, h_2(t) = (\log t)^2 \qquad \text{if } p_1 = p_2 = 0$$

Fractional polynomials provide a much wider range of shapes for curves than allowed by standard polynomials, including curves with asymptotes. However, they are still global functions.

4.3 Spline Functions

Among the limitations of polynomial functions and fractional polynomials is their global nature. Tweaking the coefficients to achieve a functional form in one region can cause the function to have a bad fit in remote regions (Hastie *et al.*, 2001). These problems can be addressed by using non parametric smoothing methods. One alternative to avoid the limitations of polynomial functions is by using splines, that is representing the function as a combination of local polynomials. Splines are functions constructed by combining pieces of polynomials. In this section we describe a piecewise polynomial representation of a function using truncated power basis functions, B-splines and M-splines.

4.3.1 Truncated Power Basis

A piecewise polynomial function $f(t)$ can be constructed by dividing the domain of t, say $[a, b]$, into intervals and fitting separate polynomial curves in each interval. The points that divide the intervals are known as knots. The function $f(t)$ is obtained by imposing continuity and differentiability on the piecewise polynomials up to a certain order at the

knots, where two adjacent segments join. For example, if the domain of t is divided by two inner knots k_1 and k_2, and three linear functions are fitted representing $f(t)$ on the three different intervals, then the function $f(t)$ is constructed based on six basis functions and hence six parameters, two for each degree 1 function. The basis functions are

$$\begin{aligned} h_0(t) &= I(t < k_1).1, & h_1(t) &= I(t < k_1).t, \\ h_2(t) &= I(k_1 \leq t < k_2).1, & h_3(t) &= I(k_1 \leq t < k_2).t, \\ h_4(t) &= I(k_2 \leq t).1, & h_5(t) &= I(k_2 \leq t).t \end{aligned}$$

Therefore, $f(t) = \sum_{l=0}^{5} \alpha_l h_l(t)$ that is,

$$f(t) = \begin{cases} \alpha_0 + \alpha_1 t, & \text{if } t < k_1 \\ \alpha_2 + \alpha_3 t, & \text{if } k_1 \leq t < k_2 \\ \alpha_4 + \alpha_5 t, & \text{if } k_2 \leq t \end{cases}$$

Let us now consider the data simulated in Section 4.1 and estimate the true function, $f(t) = sin(t) + 2$, using a piecewise linear function. Let the domain of t be divided into three intervals at knots $k_1 = 2$ and $k_2 = 6$. The fitted curve denoted by dashed lines is shown in Figure 4.2, the solid line represents the true curve. The fitted curve has three pieces of degree 1 polynomials. In the figure, it can be seen that the estimated function is discontinuous at the knots. However, since the true function is continuous throughout the domain we want the estimated function to be continuous as well. Therefore, a constraint is imposed on the parameters such that the function is continuous at the knots. Each of the fitted pieces of polynomials is constrained to be equal to a polynomial in the next interval at the knot which connects them. In the example, the constraint means $f(k_1^-) = f(k_1^+)$ and $f(k_2^-) = f(k_2^+)$. This implies that $\alpha_0 + \alpha_1 k_1 = \alpha_2 + \alpha_3 k_1$ and $\alpha_2 + \alpha_3 k_2 = \alpha_4 + \alpha_5 k_2$. These constraints reduce the number of parameters that define the piecewise linear function from 6 to 4 = (3 intervals) $\times$ (2 parameters for each interval)

- (2 knots) × (1 constraint per knot).

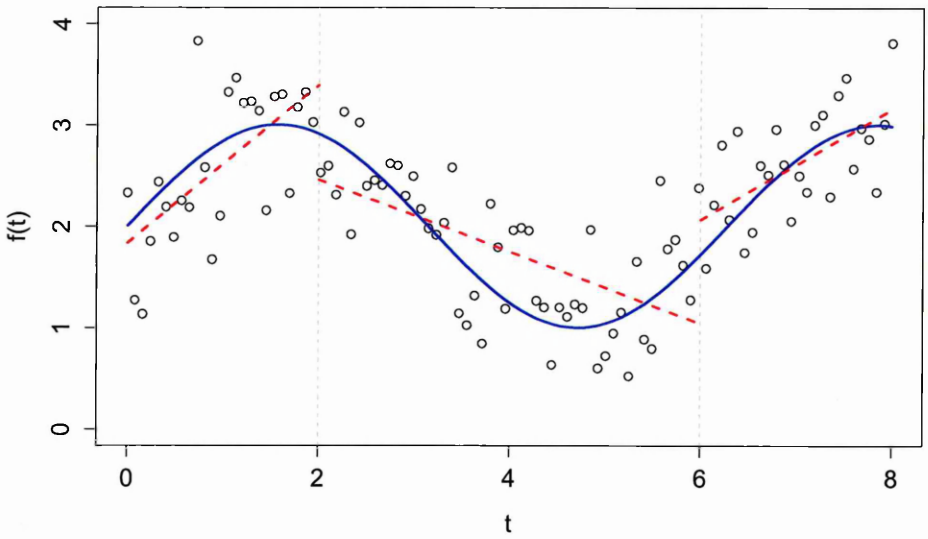

Figure 4.2: *Piecewise linear function fitted to a data simulated from a normal distribution with mean* $sin(t)+2$ *and standard deviation 0.5. The fitted curve is discontinuous and is represented by the three dashed lines and the solid line is the true function. The data are divided into three intervals at knots* 2 *and* 6

A piecewise polynomial function constrained to be continuous at the knots can be constructed by using truncated power functions as basis, which directly take the constraint into account. A truncated power function of degree p for a knot k_l is defined as:

$$(t-k_l)_+^p = (t-k)^p I_{t>k_l}(t)$$

The $+$ indicates the function takes a value 0 for t to the left of k_l and $(t-k_l)^p$ otherwise, that is

$$(t-k_l)_+^p = \begin{cases} (t-k_l)^p, & \text{if } t \geq k_l \\ 0, & \text{if } t < k_l \end{cases}.$$

The truncated power functions are used together with polynomial basis functions to form a truncated power basis. For example to fit a piecewise linear model with two knots at k_1 and k_2, the truncated power basis comprises $h_0(t) = 1$, $h_1(t) = t$, the polynomial part, and truncated power functions of degree 1, $h_3(t) = (t - k_1)_+$ and $h_4(t) = (t - k_2)_+$. Then, the piecewise linear function is estimated as:

$$f(t) = \alpha_0 + \alpha_1 t + \alpha_2(t - k_1)_+ + \alpha_3(t - k_2)_+.$$

This function, expressed as a linear combination of truncated power basis functions, is known as a spline function. From the basis functions it can be seen that if the parameters associated with the truncated power functions, α_2 and α_3, are both estimated to be zero then the function $f(t)$ reduces to a single polynomial function of degree one. Panel (a) of Figure 4.3 shows a piecewise linear function fitted to the simulated data presented in Figure 4.2 and the corresponding truncated power basis functions are shown in panel (b).

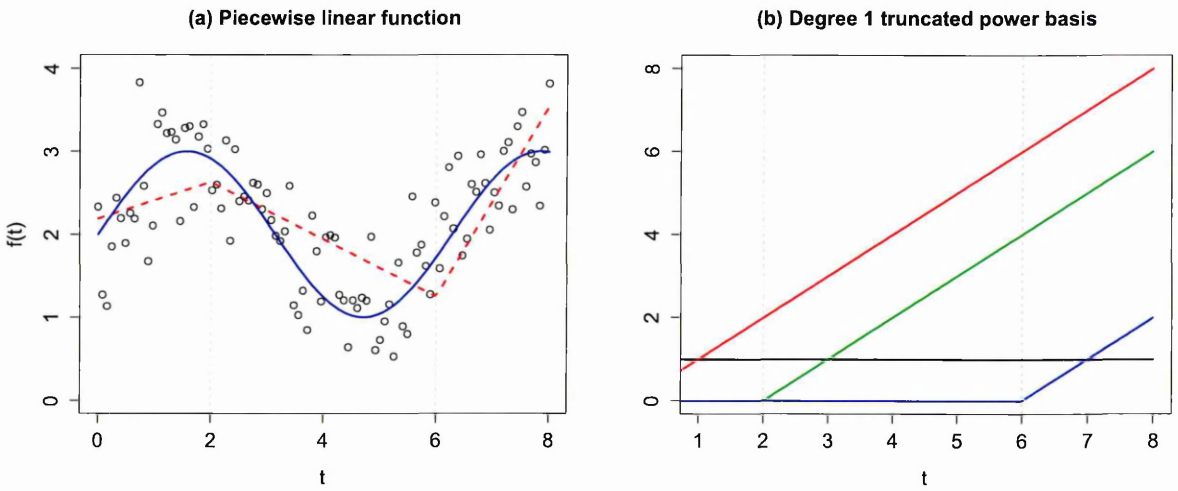

Figure 4.3: *Panel (a) piecewise linear spline function fitted to simulated data, Panel (b) Linear truncated power basis used to estimate the piecewise linear function*

It can now clearly be seen from Figure 4.3 that the use of truncated power basis leads to

a function which is continuous everywhere, linear everywhere except at the knots and has different slopes for each interval. The fitted linear spline curve in panel (a) of Figure 4.3 denoted by the dashed lines has sharp corners at the knots. This shows that the function does not have a continuous first derivative at the knots. If a smoother function is required, higher order truncated power basis functions can be employed, which are straightforward to construct. For example, a piecewise cubic polynomial function (cubic spline) with two knots has a truncated power basis that includes the degree three polynomial basis and two degree three truncated power functions, namely

$$h_0(t) = 1, h_1(t) = t, h_2(t) = t^2, h_3(t) = t^3$$

and

$$h_4(t) = (t - k_1)_+^3, h_5(t) = (t - k_2)_+^3$$

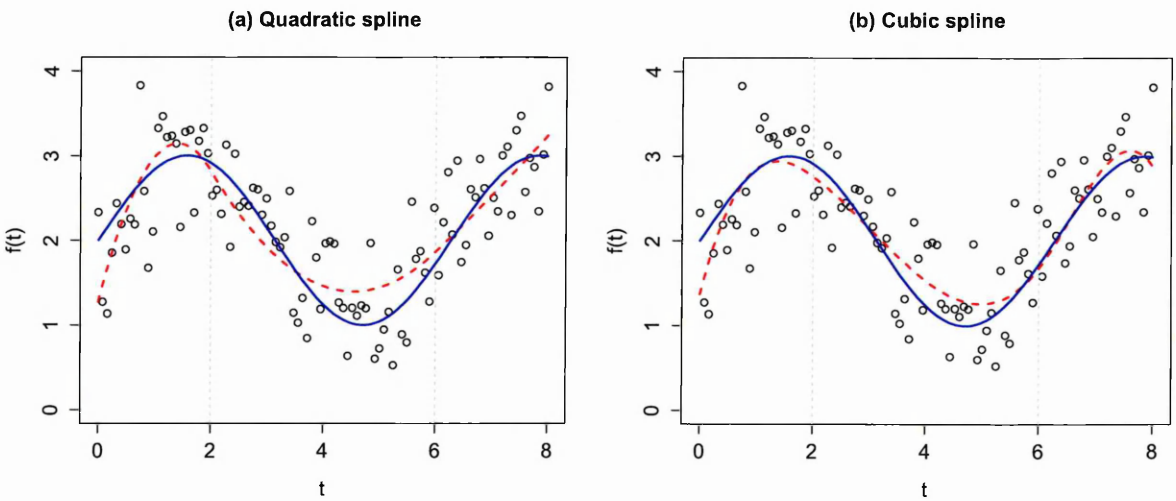

Figure 4.4: *Panel (a) piecewise quadratic spline function fitted to simulated data, Panel (b) cubic spline function. In both the panels the dashed lines represent fitted curves and the solid lines represent true curve used to simulate the data points denoted by circles*

A cubic spline has continuous first and second derivatives at the knots. Quadratic and

cubic splines fitted to the simulated data are shown on panels (a) and (b) of Figure 4.4 respectively.

In general, piecewise polynomials of order q or degree $p = q - 1$ connected at knots $k_1, k_2, ..., k_s$ have truncated power basis 1, t, ..., t^p, $(t-k_1)_+^p$, ..., $(t-k_s)_+^p$ and their linear combination gives a spline function which has continuous derivatives up to order $q-1$ as follows:

$$f(t) = \alpha_0 + \alpha_1 t + ... + \alpha_p t^p + \sum_{l=1}^{s} \alpha_{pl}(t-k_l)_+^p.$$

This function has two parts, a global polynomial of degree p and truncated power functions of degree p related to the knots. The design matrix is

$$T = \begin{bmatrix} 1 & t_1 & \dots & t_1^p & (t_1-k_1)_+^p & \dots & (t_1-k_s)_+^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & t_n & \dots & t_n^p & (t_n-k_1)_+^p & \dots & (t_n-k_s)_+^p \end{bmatrix}$$

4.3.2 B-splines

Truncated power basis functions have an advantage of simplicity to construct which makes them attractive for statistical work, they however have a rather serious disadvantage of generating considerable rounding error (Ramsay, 1988). The numerical precision problem occurs because these functions have a rapid growth without bound as t increases, especially when the domain of t is wide. Moreover, as they are far from orthogonal they can suffer from numerical instability when there is a large number of knots (Ruppert *et al.*, 2003). When computed at some value of t, many or even all of the truncated power basis functions can be non zero leading to a design matrix containing only a few zeros, which prevents the use of sparse matrix techniques to reduce computational time. These limitations of the truncated power basis functions can be avoided by using reasonably

well-conditioned and more stable equivalent basis functions known as B-splines.

B-splines are piecewise polynomial functions derived from truncated power functions and have more stable numerical properties. A B-spline of order q, $q \geq 1$, is a combination of polynomial functions of degree $q-1$ connected at knots. They are defined as appropriately scaled divided differences of truncated power functions (de Boor, 1978).

Divided differences are mathematical tools having various applications in numerical analysis including polynomial interpolation and derivation of B-spline basis functions. Let $t_0, t_1, \ldots, t_n$ be distinct real numbers and let $f(t_0), f(t_1), \ldots, f(t_n)$ be the associated function values. Then to find a polynomial function, $p(x) = \alpha_0 + \alpha_1 t + \cdots + \alpha_n t^n$, that interpolates the data, divided differences can be used. The n^{th} order divided difference of a function f is the leading coefficient of the polynomial function $p(x)$ and is denoted by $[t_0, t_1, \ldots, t_n]f$. It is evaluated as:

$$[t_0, t_1, \ldots, t_n]f = \frac{[t_1, t_2, \ldots, t_n]f - [t_0, t_1, \ldots, t_{n-1}]f}{t_n - t_0}$$

where

$$[t_0]f = f(t_0)$$

and so

$$[t_0, t_1]f = \frac{f(t_1) - f(t_0)}{t_1 - t_0}$$

etc. Now, to derive B-spline basis functions using divided differences, let the truncated power function of order q centered at t be denoted by $T_t^q(k)$

$$T_t^q(k) = (t-k)_+^{q-1}.$$

In the derivation of the B-splines, the truncated power function $(t-k)_+^{q-1}$ of the two variables t and k is taken by fixing t and considering $(t-k)_+^{q-1}$ as a function of k only (de Boor, 1978).

Let the function to be approximated by a spline be defined in the interval $[a, b]$ and $k_1 = k_2 = \cdots = k_q < k_{q+1} < \cdots < k_{q+s} < k_{q+s+1} = k_{q+s+2} = \cdots = k_{2q+s}$, be a non-decreasing sequence of knots in $[a, b]$, where $k_q = a$ and $k_{q+s+1} = b$. s is the number of interior knots, $k_{q+1} < \cdots < k_{q+s}$, which are within the domain $[a, b]$ and to create the B-spline basis functions, the first and last $q - 1$ knots are arbitrarily added. Usually the extra knots at the beginning of the sequence are all taken to be equal to a and the ones at the end equal to b. But it is also possible to give them arbitrary values. Then, the l^{th} B-spline of order q at t for the given knot sequence denoted by $B_l(t|q)$ is defined as an appropriately scaled q^{th} order divided difference at $k_l, k_{l+1}, \ldots, k_{l+q}$ of the truncated power function $T_t^q(k)$ (de Boor, 1978), by the rule

$$B_l(t|q) = (-1)^q(k_{l+q} - k_l)[k_l, \ldots, k_{l+q}]T_t^q(k)$$

for $l = 1, 2, ..., m$ and $m = q + s$ is the number of B-spline basis functions, which is equal to the number of interior knots plus the order q. $[k_l, \ldots, k_{l+q}]T_t^q(k)$ is the q^{th} divided difference at $k_l, \ldots, k_{l+q}$ of $T_t^q(k)$. Here $k_{l+q} > k_l$, if $k_{l+q} = k_l$ then $B_l(t|q)$ is defined to be zero.

Each $B_l(t|q)$ is positive over the interval $k_l \leq t < k_{l+q}$ and zero elsewhere, i.e it is non zero over q intervals in the domain of t, $[a, b]$, and each interval has q positive B-splines.

The q^{th} divided difference of the truncated power function, in the definition of B-splines, is multiplied by $(k_{l+q} - k_l)(-1)^q$ so that at a given value of t the sum of the q positive B-splines is equal to one, i.e $\sum_l B_l(t|q) = 1$. Each $B_l(t|q)$ consists of q polynomial pieces of degree $q - 1$ that are joined at $q - 1$ inner knots and whose derivatives up to order $q - 2$ are continuous at the joining points. The other property of B-splines is that the integral of each of $B_l(t|q)$ between k_l and k_{l+q} is $\int_{k_l}^{k_l+q} B_l(t|q)dt = \frac{k_{l+q}-k_l}{q}$.

As an example let us consider evaluating B-splines of order 2 ($q = 2$) for the data simulated in Section 4.1 where the values of t range from 0 to 8. Let the inner knots be 2, 4 and 6, then the set of knots including the minimum and maximum values of t will be $\{0, 2, 4, 6, 8\}$. To create a B-spline we add an arbitrary extra $q - 1 = 1$ knots at the beginning and end of the set, which gives a total of 7 knots $\{0, 0, 2, 4, 6, 8, 8\}$. Then, for example, the 3^{rd} B-spline $B_3(t|2)$ ($l = 3$) is

$$\begin{aligned} B_3(t|2) &= (-1)^2(k_{3+2} - k_3)[k_3, k_4, k_5]T_t^2(k) \\ &= (k_5 - k_3)\frac{[k_4, k_5]T_t^2(k) - [k_3, k_4]T_t^2(k)}{(k_5 - k_3)} \\ &= \frac{T_t^2(k_5) - T_t^2(k_4)}{k_5 - k_4} - \frac{T_t^2(k_4) - T_t^2(k_3)}{k_4 - k_3} \\ &= \frac{(t-6)_+ - (t-4)_+}{6-4} - \frac{(t-4)_+ - (t-2)_+}{4-2} \end{aligned}$$

For $t < k_3 = 2$

$$\begin{aligned} B_3(t|2) &= \frac{(0) - (0)}{2} - \frac{(0) - (0)}{2} \\ &= \frac{0}{2} - \frac{0}{2} = 0 \end{aligned}$$

For $k_3 = 2 \leq t < k_4 = 4$

$$\begin{aligned} B_3(t|2) &= \frac{(0) - (0)}{2} - \frac{(0) - (t-2)}{2} \\ &= \frac{t-2}{2} \end{aligned}$$

For $k_4 = 4 \leq t < k_5 = 6$

$$\begin{aligned} B_3(t|2) &= \frac{(0) - (t-4)}{2} - \frac{(t-4) - (t-2)}{2} \\ &= \frac{6-t}{2} \end{aligned}$$

And for $t \geq k_5 = 6$

$$\begin{aligned} B_3(t|2) &= \frac{(t-6)-(t-4)}{2} - \frac{(t-4)-(t-2)}{2} \\ &= \frac{-2}{2} - \frac{-2}{2} = 0 \end{aligned}$$

Since the number of inner knots is 3, and the order is 2 the total number of basis functions is 5 and the remaining B-spline functions, $B_1(t|2), B_2(t|2)$, $B_4(t|2)$, and $B_4(t|2)$ can be obtained in a similar way.

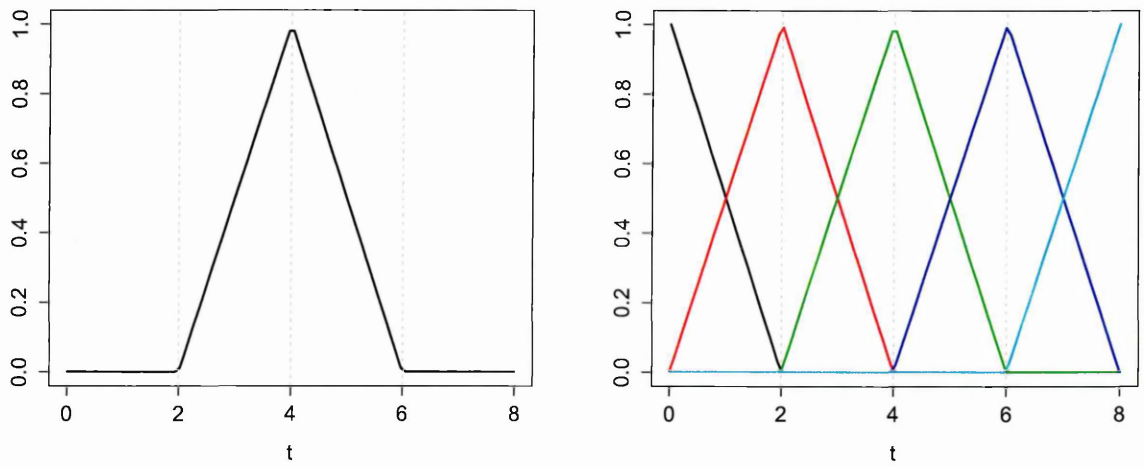

Figure 4.5: *B-spline basis functions of order 2: Left panel one basis function which is a combination of two linear functions and the right panel all the five B-spline basis functions*

Graphically, $B_3(t|2)$, is a triangular hat function with a value of zero outside the interval between 2 and 6 and is presented in the left panel of Figure 4.5. The right panel shows all the five basis functions, the vertical dotted lines are the values of t where the knots are located. The data used to plot these basis functions are those simulated in Section 4.1. The first and last basis functions are discontinuous at the minimum and maximum values of t because there are repeated knots at those locations. Order two

spline function fitted from B-spline basis functions of order two is shown in Panel *a* of Figure 4.7.

B-spline basis functions of order 4 are known as cubic B-splines because they are linear combinations of degree 3 polynomials, and can be obtained as explained for order 2 B-splines. Figure 4.6 shows cubic B-spline basis functions. The inner knots are at 2, 4, and 6 with minimum and maximum values of t at 0 and 8 respectively. The left panel of Figure 4.6 shows the fourth B-spline, it covers four intervals and each interval is a piece of a degree three polynomial. And the right panel shows all the seven basis functions. A spline function fitted from these basis functions is presented in panel *b* of Figure 4.7

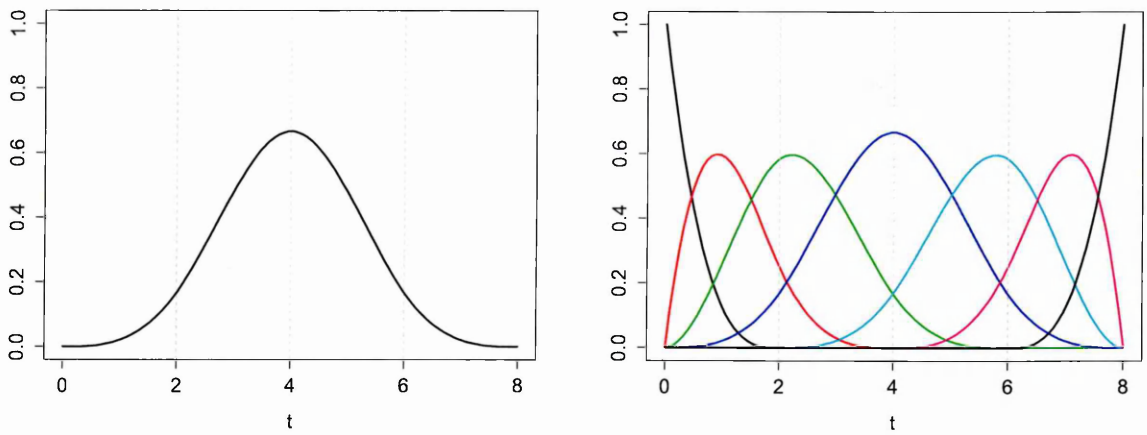

Figure 4.6: *B-splines of order four: Left panel one basis function which is a combination of four cubic polynomial pieces and the right panel all the seven B-spline basis functions*

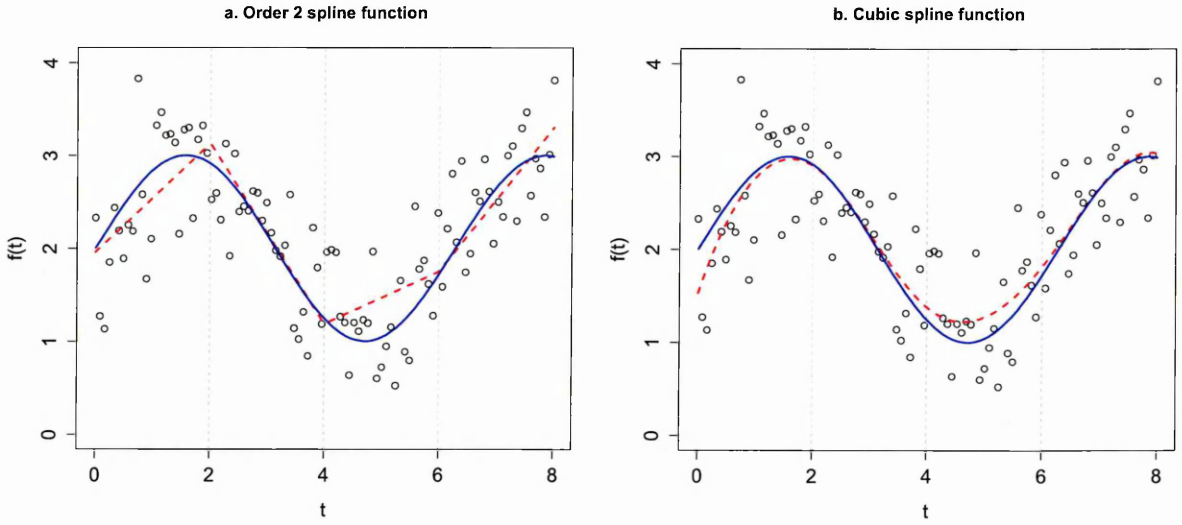

Figure 4.7: *Spline functions fitted using B-spline basis functions. In panel a B-splines of order two are used and in panel b cubic B-splines*

From the computational perspective B-splines can be obtained using recurrence relations given by the Cox-de Boor algorithm (de Boor, 1978), instead of being obtained directly from the truncated power functions. They are computed from B-splines of lower order. It is simple to compute B-spline of any order since the B-spline of order one is constant between two knots. Computing B-splines from the truncated power functions directly may lead to the same problem as using a truncated power basis (de Boor, 1978). The recursion algorithm was developed on the basis of the Leibniz formula for the q^{th} order divided difference of a product, which states that, if a function $f(t) = g(t)h(t)$ then the q^{th} order divided difference of $f(t)$ at $k_l, \ldots, k_{l+q}$ is given by

$$[k_l, \ldots, k_{l+q}]f = \sum_{r=l}^{l+q} ([k_l, \ldots, k_r]g)([k_r, \ldots, k_{l+q}]h)$$

The Leibniz formula is applied to the truncated power function $T_t^q(k) = (t-k)_+^{q-1}$ by first

expressing it as a product two functions as

$$(t-k)_+^{q-1} = (t-k)(t-k)_+^{q-2}$$

which leads to the following expression for a B-spline of order q:

$$B_l(t|q) = \begin{cases} \frac{t-k_l}{(k_{l+q-1}-k_l)} B_l(t|q-1) + \frac{k_{l+q}-t}{k_{l+q}-k_{l+1}} B_{l+1}(t|q-1), & k_l \leq t < k_{l+q} \\ 0, & \text{elsewhere,} \end{cases}$$

with

$$B_l(t|1) = \begin{cases} 1, & k_l \leq t < k_{l+1} \\ 0, & \text{elsewhere.} \end{cases}$$

Once the B-spline basis functions are computed, their linear combination gives the desired spline function. The design matrix, denoting $B_l(t|q)$ as $B_l(t)$ and m representing the total number of B-spline basis functions (number of interior knots + order of the spline), is:

$$T = \begin{bmatrix} B_1(t_1) & B_2(t_1) & \dots & B_m(t_1) \\ \vdots & \vdots & \vdots & \vdots \\ B_1(t_n) & B_2(t_n) & \dots & B_m(t_n) \end{bmatrix}$$

Then the spline function is defined as

$$f(t) = \sum_{l=1}^{m} \alpha_l B_l(t).$$

The design matrix from B-splines contains few non zero elements unlike the one obtained from the truncated power basis, because a given order q B-spline basis is non zero only in q intervals in the domain of t. This leads to efficient computation.

Figure 4.7 shows spline functions fitted, using B-spline basis functions of order two (Panel a) and order four (Panel b), to a data set of 100 data points simulated from a gaussian distribution with mean $sin(t) + 2$ and standard deviation of 0.5. The values of t

range from 0 to 8. The fitted curves are denoted by dashed lines, the solid lines represent the true curve and the circles represent the data points.

4.3.3 M-splines

M-splines, which are variants of B-splines, were first introduced by Curry and Schoenberg (1947). They use a different method of normalization from the B-splines discussed in the previous section. They are normalized such that $\int_{-\infty}^{\infty} M_l(t|q) = 1$. The M-splines, like B-splines, have the properties that:

1. $M_l(t|q) = 0$ outside $k_l \leq t < k_{l+q}$,
2. they are positive functions in the interval $k_l \leq t < k_{l+q}$,
3. each M-spline of order q is a linear combination of q polynomial pieces of degree $q-1$,
4. $M_l(t|q)$ has $q-2$ continuous derivatives at the knots,
5. M-splines are related to B-splines as, $M_l(t|q) = (\frac{q}{(k_{l+q}-k_l)})B_l(t|q)$.

Similar to B-splines, M-splines can be computed directly as a divided difference of truncated power functions or using the recursion algorithm of de Boor (1978). Given a knot sequence $k_1 = k_2 = \cdots = k_q < k_{q+1} < \cdots < k_{q+s} < k_{q+s+1} = k_{q+s+2} = \cdots = k_{2q+s}$, an M-spline of order q is defined as

$$M_l(t|q) = \begin{cases} \frac{q[(t-k_l)M_l(t|q-1)+(k_{l+q}-t)M_{l+1}(t|q-1)]}{(q-1)(k_{l+q}-k_l)}, & k_l \leq t < k_{l+q} \\ 0, & \text{elsewhere,} \end{cases}$$

with

$$M_l(t|1) = \begin{cases} \frac{1}{(k_{l+1}-k_l)}, & k_l \leq t < k_{l+1} \\ 0, & \text{elsewhere.} \end{cases}$$

The spline function can now be obtained as a linear combination of M-splines,

$$f(t) = \sum_{l=1}^{m} \alpha_l M_l(t|q).$$

Since M-splines are positive functions, their linear combination, $\sum_{l=1}^{m} \alpha_l M_l(t|q)$, can be used to approximate a non-negative function by constraining their coefficients to be non-negative, $\alpha_l \geq 0$. $M_l(t|q)$ is zero outside $[k_l, k_{l+q}]$ and positive inside this interval, therefore any change in the coefficient α_l has an effect only in the interval hence local sensitivity to coefficient changes. The integral of $f(t)$ can be constrained to be equal to one, $\int_a^b \sum_{l=1}^{m} \alpha_l M_l(t|q) = 1$, by setting the sum of the coefficients to be one, $\sum_{l=1}^{m} \alpha_l = 1$, since $\int_{k_l}^{k_{l+q}} M_l(t|q) = 1$. In addition, as only M_1 and M_m are non zero at $f(a)$ and $f(b)$ respectively, $f(a) = 0$ or $f(b) = 0$ can be obtained by setting $\alpha_1 = 0$ or $\alpha_m = 0$ respectively (Ramsay, 1988). Since each $M_l(t|q)$ is a piecewise polynomial, linearity, differentiability and integrability properties of polynomials carryover to the spline function.

Cubic M-splines of order four with three interior knots are shown in the left panel of Figure 4.8. The shape of the M-splines are similar to the B-splines in Figure 4.6 but are different in their values at a given t. For B-splines the maximum value is one because they are normalized such that the sum of all B-splines at a given value of t is one however, M-splines can have a value greater than one.

4.3.4 I-splines

Ramsay (1988) defined integrated splines (I-splines) to be used as basis functions in regression analysis when monotonicity is required. I-splines are piecewise polynomials of degree q obtained by integrating M-splines of degree $q-1$ and are thus defined for $k_h \leq t < k_{h+1}$ as $I_l(t|q) = \int_a^t M_l(u|q)du$.

Thus for the same sequence of interior knots used in M-splines, I-splines are defined

as

$$I_l(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h}(k_{m+q+1} - k_m)\frac{M_m(t|q+1)}{q+1}, & h-q+1 \leq l \leq h \\ 1 & l < h-q+1. \end{cases}$$

One of the properties of M-splines is that $\int_{k_l}^{k_{l+q}} M_l(t|q) = 1$, therefore, I-splines are monotone splines constrained between 0 and 1. A linear combination of I-splines is a monotone spline function if the coefficients α_l are constrained to be non-negative. I-splines of order five can be seen graphically in the right panel of Figure 4.8.

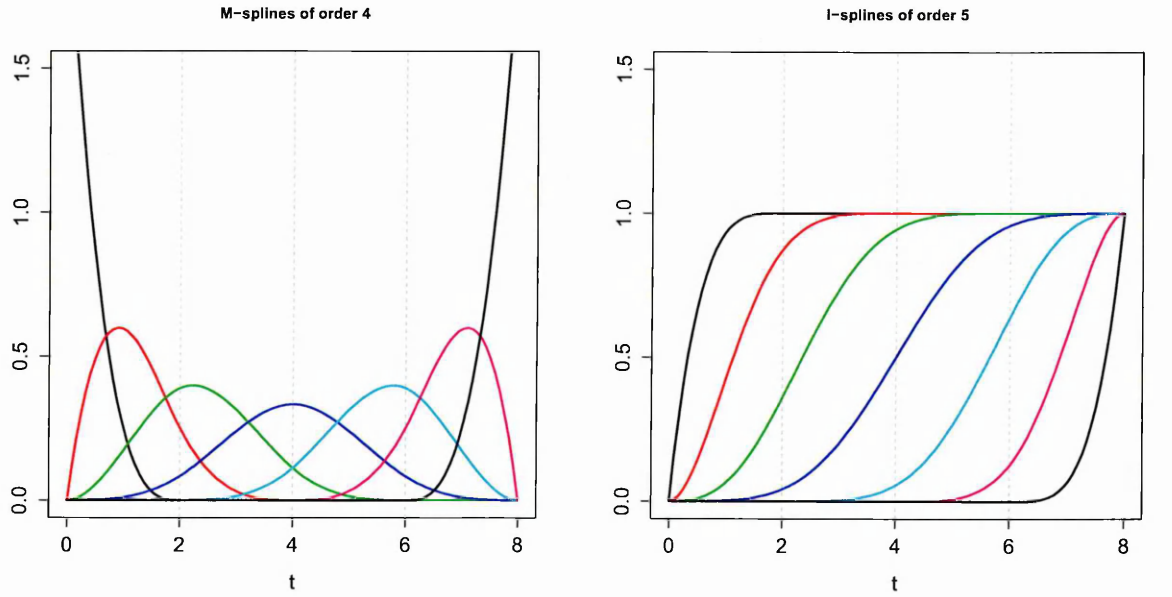

Figure 4.8: *Left panel: M-spline basis functions of order 4. Right panel: I-splines of order 5 obtained from the M-splines in the left panel.*

The maximum value of an I-spline $I_l(t|q)$ is one when the value of t is greater or equal to k_{l+q}.

4.4 Discussion

In this chapter, we described methods of approximating a curve with smooth functions that could be employed to model age and exposure effects in the self-controlled case series method. The parametric models presented (polynomial and fractional polynomials) are flexible ways of modelling, however they fail to follow deviations from the overall trend of the data. If one chooses a parametric model that is not of appropriate form, at least approximately, then there is a danger of reaching incorrect conclusions (Wand and Jones, 1995). The restrictive nature of polynomials can be avoided by using non-parametric smoothing methods.

Piecewise polynomial models or splines are one class of smoothing methods which can be used to allow the data to decide the shape of the estimated function. The spline function is estimated by imposing continuity and differentiability conditions up to a certain order at the knots, where the pieces of polynomials are joined. The choice of knots is crucial in using splines. We demonstrated, in the previous sections, the use of different type of splines based on a priori chosen fixed knots. These types of splines are known as regression splines. Too large a number of knots can over-fit the data resulting in a rough function and too low a number of knots leads to under-fitting. There are several ways of defining the number and location of knots. Methods to automatically choose knots have been proposed in the literature including Friedman (1991); Smith and Kohn (1996); Dimatteo *et al.* (2001), however these automatic knot selection procedures are quite complicated and computationally intensive (Ruppert *et al.*, 2003). So another approach is to choose a large number of knots and constrain their influence by introducing a penalty as proposed by O'Sullivan (1986). The penalty term proposed by O'Sullivan

(1986) is based on the second derivative of the fitted spline function, $\lambda \int \{f''(t)\}^2$. λ is known as a smoothing parameter that controls the tradeoff between smoothness and fit. Following O'Sullivan (1986) different types of penalties have been proposed in the literature including Eilers and Marx (1996), who apply a difference penalty directly to adjacent coefficients, and Ruppert *et al.* (2003) who make use of a ridge penalty as in ridge regression. The locations of the knots are usually chosen to be equidistant or selected based on quantiles. Smoothing splines, which are different from regressions splines, avoid the problem of choosing knots by using all distinct data points as knots, and to prevent over fitting a penalty similar to O'Sullivan (1986) is used. The use of smoothing splines is computationally demanding because the number of parameters to be estimated is equal to about the number of distinct number of observations.

When using regression splines, in addition to the choice of knots, the choice of basis functions, the order of the basis functions, continuity constraints on the function to be estimated, the penalty method and the method of choosing the smoothing parameter are important. The most commonly used basis functions as described in this chapter are polynomials, truncated power basis functions, B-splines, M-splines. Other types of basis functions include radial basis functions, and the convex spline basis (C-splines).

An alternative to splines which can be used to model time-varying covariates in SCCS is kernel smoothing (Wand and Jones, 1995). One of the advantages of kernel smoothers as compared to splines is their simple theoretical analysis (Ruppert *et al.*, 2003). Similar to splines, kernel smoothers are local, so are able to follow deviations from the overall trend in the data unlike polynomials. Splines and kernel smoothers are obtained by fitting local polynomials but the way they are estimated is different. The kernel smoothers are estimated by fitting local polynomials at each data point based on neighbouring

data points having different weights. Closer values are given higher weights, therefore have higher influence on the local polynomial fit at a point. But spline functions are estimated by fitting pieces of polynomials in different intervals and constraining them to be continuous at the knots where the intervals are joined.

In this thesis we will use cubic M-splines and I-splines as basis functions in modelling age and exposure effects for the self-controlled case series models. As noted by Ramsay and Silverman (1997), we want the basis functions to allow constrained modelling, namely a positivity condition. In the self-controlled case series models the functions related to age-specific relative incidence and exposure relative incidence should be positive functions. Therefore, using M-splines as a basis enables one to fit a non-negative function by constraining their coefficients to be non-negative, since M-splines are non-negative functions. In addition, the use of M-splines avoids numerical integration of the denominator in the log-likelihood function of the SCCS method because integrals of M-splines can be expressed as other forms of basis functions known as I-splines.

Chapter 5

Smooth Age Effect

For a self-controlled case series analysis, time dependent variables (e.g age), unlike fixed covariates, are not automatically controlled for, and therefore need to be included in the model. Confounding by temporal factors is likely to occur when both the event incidence and the opportunity for exposure vary with age or season. Examples include adverse events and childhood vaccinations; seasonal exposures such as respiratory infections or influenza vaccination; and studies in elderly populations of age-related conditions. Therefore a careful control of age is important.

As discussed in Section 2.3 of Chapter 2, the methods for handling an age effect in the standard and semi-parametric versions of SCCS method have limitations. The standard method can be sensitive to mis-specification of age groups which may lead to biased estimates of the association between exposure and event outcome. The semi-parametric method may run into computational problems when the number of cases is moderately large and there may also be a loss of efficiency in estimation. In this chapter, we propose the use of smooth functions to represent the age-specific relative incidence function to avoid these limitations. We consider polynomial, fractional polynomial functions and a

linear combination of M-spline functions.

The use of M-spline functions, in addition to addressing the above mentioned limitations of the standard and the semi-parametric SCCS methods, avoids the integral in the case series likelihood function by replacing it with I-splines, which are integrated M-splines. The age-specific cumulative relative incidence function is then approximated by a monotone spline function, a linear combination of I-splines (Ramsay, 1988). A penalised log-likelihood approach is used in estimating the parameters related to the spline function. The methodology developed here is inspired by Joly *et al.* (1998), which we adapted for use with the SCCS method. Our methods have been programmed in R (R Development Core Team, 2012).

The outline of this Chapter is as follows. Section 5.1 describes how to fit an SCCS model with the log of the age-specific relative incidence function represented by a polynomial function and reviews a recent paper by Lee and Carlin (2014) that used fractional polynomials to estimate parameters in the standard SCCS method. This is followed by a description of an M and I-splines based SCCS model in Section 5.2. In Section 5.3, a simulation study that evaluates the performance of the spline-based SCCS and its comparison with the standard and semi-parametric versions of SCCS is presented. In Section 5.4, we apply the new spline-based SCCS to a large data set on febrile convulsions and paediatric vaccines, and some final remarks are made in Section 5.5. The contribution of this chapter has been published in Ghebremichael-Weldeselassie *et al.* (2014a).

5.1 Modelling of Age Effect Using Parametric Functions

In this section, we consider modelling of the age effect with smooth parametric functions, specifically polynomial and fractional polynomial functions. We also discuss the limitations of using such functions.

5.1.1 Polynomial Functions

As a first step, we replace the log of the age-specific relative incidence function, $\psi(t)$, in the SCCS likelihood function (2.9) by a smooth polynomial function. That is, $\psi(t)$ will be the exponential of a polynomial function, while the relative incidence associated with exposure, $x_i(t)$, remains as a parametric step function. The SCCS log-likelihood function, when the log of the age-specific relative incidence is represented by a polynomial function of order 2, is derived as follows. Suppose each individual i has just one exposure period $(c_i, d_i]$ as shown in Figure 5.1

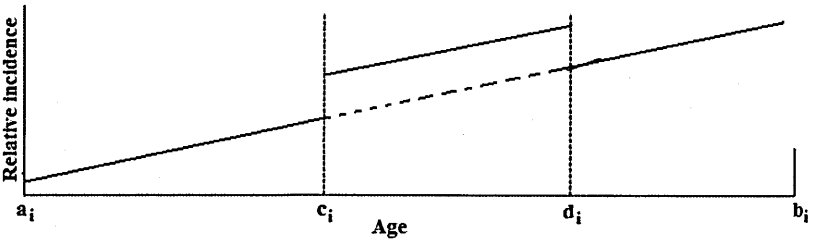

Figure 5.1: *Relative risk for individual i in different periods within the observation period when the log of age-specific relative incidence function is represented by a linear function*

In Figure 5.1 c_i and d_i represent the start and end of the exposure risk period for individual i, while a_i and b_i represent the beginning and end of the observation period. In

the first interval, a_i to c_i, and third interval, d_i to b_i, the relative incidence rate is simply $\exp(\alpha_0 + \alpha_1 t)$, which is related only to age because there is no effect of exposure in those periods. Whereas, the incidence rate in the exposure period c_i to d_i is $\exp(\alpha_0 + \alpha_1 t) \times \exp(\beta)$, where $\exp(\beta)$ is the relative incidence associated with exposure and $log(\psi(t)) = \alpha_0 + \alpha_1 t$. The likelihood function for individual i will then be

$$
\begin{aligned}
L &= \prod_{i=1}^{N}\prod_{j=1}^{n_i} \frac{\exp(\alpha_0 + \alpha_1 t_{ij}) \exp\{x_i(t_{ij})\beta\}}{\int_{a_i}^{c_i} \exp(\alpha_0 + \alpha_1 t)dt + \int_{c_i}^{d_i} \exp(\alpha_0 + \alpha_1 t)\exp(\beta)dt + \int_{d_i}^{b_i} \exp(\alpha_0 + \alpha_1 t)dt} \\
&= \prod_{i=1}^{N}\prod_{j=1}^{n_i} \frac{(\alpha_1)\exp(\alpha_1 t_{ij}) \exp\{x_i(t_{ij})\beta\}}{\exp(\alpha_1 c_i) - \exp(\alpha_1 a_i) + \exp(\beta)\left(\exp(\alpha_1 d_i) - \exp(\alpha_1 c_i)\right) + \exp(\alpha_1 b_i) - \exp(\alpha_1 d_i)}
\end{aligned}
$$

where $x_i(t) = 1$ if individual i is exposed at t and 0 otherwise. Note that α_0 the intercept of the linear function that represents log of the age effect cancels out. The log-likelihood function is then

$$
\begin{aligned}
l &= \sum_{i=1}^{N}\sum_{i=1}^{n_i} \alpha_1 t_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{n_i} x_i(t_{ij})\beta \\
&\quad - \sum_{i=1}^{N}\sum_{i=1}^{n_i} \log\left(\frac{1}{\alpha_1}\left(\exp(\alpha_1 c_i) - exp(\alpha_1 a_i) + \exp(\beta)(\exp(\alpha_1 d_i) - \exp(\alpha_1 c_i)) + \right.\right. \\
&\qquad \left.\left. \exp(\alpha_1 b_i) - \exp(\alpha_1 d_i)\right)\right) \qquad (5.1)
\end{aligned}
$$

The parameters of interest are obtained by maximising the log-likelihood function (5.1). The log-likelihood can be derived in the same way for polynomials of higher order. However the integrals in the denominator can not be integrated analytically and have to be integrated numerically which makes estimation computationally costly.

5.1.2 Fractional Polynomials

A family of fractional polynomials, as mentioned in Chapter 4, were introduced by Royston and Altman (1994) to circumvent the limitations of the conventional polynomial functions. Fractional polynomials use integers and fractions as powers. Usually fractional polynomials of order less than or equal to two (three parameters) are

sufficient, where their powers are chosen from a set of numbers between -2 and 3, $\{-2, -1, -0.5, 0, 0.5, 1, 2, ..., \max(3, m)\}$.

In the context of the self-controlled case series method, Lee and Carlin (2014) used fractional polynomials to estimate parameters related to the age effect in the standard SCCS method. To apply the method of Lee and Carlin (2014), the data still need to be expanded as for the parametric SCCS and the effect of age is estimated by only two parameters related to fractional polynomials. Let the observation period of individual i be divided into six age groups as presented in Figure 5.2.

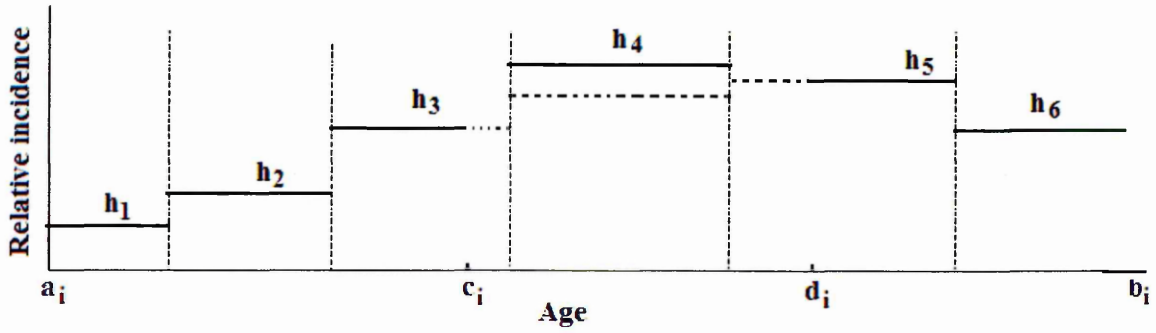

Figure 5.2: *Age groups used in estimating the age-specific relative incidence function using fractional polynomials*

h_1, h_2, h_3, h_4, h_5 and h_6 in Figure 5.2 show the age groups. Using the standard SCCS method 5 parameters are needed to estimate the age-specific relative incidence function. The method of Lee and Carlin (2014) needs only one or two parameters of fractional polynomials to estimate the jumps of the step function in Figure 5.2, by first specifying the age groups. It is also possible to use each day within the observation period as an age group, hence expand the data to size $N \times T$ where T is the number of days in an observation period. However, instead of expanding the data to use fractional polynomials,

it is possible to use the same procedure as for a polynomial function. The log-likelihood function, with log of age-specific relative incidence represented by a fractional polynomial function, can then be derived in the same way as for a polynomial function presented in Section 5.1.1, but will also require numerical integration.

Unlike the standard and semi-parametric SCCS models the use of polynomial or fractional polynomials to represent the age effect does not require the data to be expanded. The size of the data is only of order N, where N is the total number of cases and the number of parameters estimated is much lower than for the standard and semi-parametric SCCS methods. However, there is the integral involved in the denominator of the SCCS likelihood function which might make computation difficult when the polynomial is of high order. This might be simplified by choosing a polynomial or fractional polynomial that represents the age-specific relative incidence function, $\psi(t)$, directly and not the log of it. This is possible since the basis functions for the age variable, when using the polynomials or fractional polynomials, are always positive. To make the linear combination of the basis functions, combined linearly to form the age-specific relative incidence function, positive, their coefficients can be constrained to be non-negative.

5.2 Modelling of the Age Effect Using M-spline Functions

The parametric models described in the previous section, polynomial and fractional polynomial functions are flexible. However, they are a priori defined shapes through their specific analytical form (Hens *et al.*, 2012). Moreover, if a function to be approximated is badly behaved anywhere in the interval of approximation, then the approximation is

poor everywhere (de Boor, 1978). The use of piecewise polynomial functions (splines) offer flexibility without the same limitations.

One possibility is to use truncated power basis functions, discussed in Section 4.3.1, to approximate $\psi(t)$. A linear combination of these basis functions gives a spline function that represents the age effect in the SCCS method. The use of splines avoids predefining age groups and there are few computational problems involved in estimating the parameters of interest as $\psi(t)$ can be represented with few parameters. However, the use of truncated power functions as a basis has the disadvantage of numerical instability when the number of knots is large (Ruppert *et al.*, 2003). M-splines, which are derived from truncated power basis functions are more stable numerically. Since each of M_l, where M_l is the l^{th} M-spline basis function, is zero outside an interval $[k_l, k_{l+q}]$, any bad approximation within the interval does not affect the other parts of the function to be approximated, unlike polynomial functions.

We use a linear combination of M-spline basis functions to approximate $\psi(t)$ following Joly *et al.* (1998), who used these functions to approximate the hazard function in arbitrarily censored and truncated data, with an application to the age-specific incidence of dementia. In the SCCS setting, if a_i and b_i are as defined in Chapter 2, the start and end of the observation period for each individual i respectively, the spline function is defined between a and b where $a = \min\{a_i; i = 1, \ldots, N\}$ and $b = \max\{b_i; i = 1, \ldots, N\}$ respectively. So the interval [a, b] spans all the observation periods. The interior knots selected to define the M-spline basis functions should therefore be between a and b. All the required knots will then be defined by repeating the values of a and b the order of the M-splines, q, times or adding equidistant $q - 1$ knots below a and above b in addition to the interior knots chosen. Since it is a relative hazard, the age-specific relative incidence

function has to be a positive function. In this respect, M-splines are very useful as they are positive functions and to keep their linear combination non-negative, their coefficients are constrained to be non-negative. Therefore, our approximation of the age-specific relative incidence function is a linear combination of cubic M-splines, namely M-spline basis functions of order 4, and is given by:

$$\psi(t) = \sum_{l=1}^{m} g(\alpha_l) M_l(t) \tag{5.2}$$

where m is the number of basis functions, the coefficients $g(\alpha_l)$ are parameters estimated to determine the shape of the function. The non-negativity of the function $\psi(t)$ is achieved by constraining the coefficients to be non-negative. We use $g(\alpha_l) = \alpha_l^2$, hence $\psi(t) = \sum_{l=1}^{m} \alpha_l^2 M_l(t)$. $g(\alpha_l) = \exp(\alpha_l)$ can also be used but it may have a convergence problem when $g(\alpha_l)$ should be zero. Combining (2.9) and (5.2) we obtain the log-likelihood function for the SCCS model as:

$$l = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log \left(\frac{(\sum_{l=1}^{m} \alpha_l^2 M_l(t_{ij})) \exp\{x_i(t_{ij})^T \beta\}}{\int_{a_i}^{b_i} (\sum_{l=1}^{m} \alpha_l^2 M_l(t)) \exp\{x_i(t)^T \beta\}\, dt} \right). \tag{5.3}$$

A motivating reason for using M-splines to approximate the age-specific relative incidence function is that the log-likelihood function contains integrals. These can be replaced by other spline basis functions known as I-splines without the need for numerical integration since I-splines are integrals of M-splines. For example, suppose that there is only one exposure period $(c_i, d_i]$ for each individual i. Thus, $x_i(t)$ is 1 if t is in the interval $(c_i, d_i]$ and 0 otherwise. This yields the log-likelihood

$$l = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log \left(\frac{(\sum_{l=1}^{m} \alpha_l^2 M_l(t_{ij})) \exp\{x_i(t_{ij})\beta\}}{\int_{a_i}^{c_i} (\sum_{l=1}^{m} \alpha_l^2 M_l(t))dt + \exp(\beta) \int_{c_i}^{d_i} (\sum_{l=1}^{m} \alpha_l^2 M_l(t))dt + \int_{d_i}^{b_i} (\sum_{l=1}^{m} \alpha_l^2 M_l(t))dt} \right)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log \left(\frac{(\sum_{l=1}^{m} \alpha_l^2 M_l(t_{ij})) \exp\{x_i(t_{ij})\beta\}}{\int_{a_i}^{b_i} (\sum_{l=1}^{m} \alpha_l^2 M_l(t))dt - \int_{c_i}^{d_i} (\sum_{l=1}^{m} \alpha_l^2 M_l(t))dt + \exp(\beta) \int_{c_i}^{d_i} (\sum_{l=1}^{m} \alpha_l^2 M_l(t))dt} \right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{(\sum_{l=1}^{m}\alpha_l^2 M_l(t_{ij}))\exp\{x_i(t_{ij})\beta\}}{\sum_{l=1}^{m}\alpha_l^2\int_{a_i}^{b_i} M_l(t)dt + (\exp(\beta)-1)(\sum_{l=1}^{m}\alpha_l^2\int_{c_i}^{d_i} M_l(t)dt)}\right) \tag{5.4}$$

then as $I_l(t) = \int_a^t M_l(u)du$, replacing the integral of M-splines with I-splines provides the log-likelihood

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{(\sum_{l=1}^{m}\alpha_l^2 M_l(t_{ij}))\exp\{x_i(t_{ij})\beta\}}{I(b_i) - I(a_i) + (\exp(\beta)-1)(I(d_i) - I(c_i))}\right), \tag{5.5}$$

where $I(y) = \sum_{l=1}^{m}\alpha_l^2 I_l(y)$

5.2.1 Penalised Log-likelihood

As described in Chapter 4, approximating functions with splines requires determining the number and position of knots in advance. We use a large number of knots that over-fit the function and use a penalty that controls the balance between data fit and roughness. Following O'Sullivan (1986), we use a penalty function that measures roughness as the integrated squared second derivative, $\int_a^b \left(\psi''(t)\right)^2 dt$, and define our penalised log-likelihood as

$$pl = l - \lambda\int_a^b \left(\sum_{l=1}^{m} g(\alpha_l)M_l''(u)\right)^2 du$$

with $\psi''(t) = \sum_{l=1}^{m} g(\alpha_l)M_l''(t) = g(\boldsymbol{\alpha})\boldsymbol{M}''(t)$, where $\boldsymbol{M}''(t) = [M_1''(t), M_2''(t), \ldots M_l''(t)]^T$ and $g(\boldsymbol{\alpha}) = [g(\alpha_1), g(\alpha_2), \ldots g(\alpha_l)]$. Therefore,

$$\begin{aligned} pl &= l - \lambda\int_a^b g(\boldsymbol{\alpha})^T \boldsymbol{M}^{T''}(t)\boldsymbol{M}''(t)g(\boldsymbol{\alpha})dt \\ &= l - \lambda g(\boldsymbol{\alpha})^T\int_a^b \boldsymbol{M}^{T''}(t)\boldsymbol{M}''(t)dt g(\boldsymbol{\alpha}) \\ &= l - \lambda((g(\boldsymbol{\alpha}))^T\mathbf{A}g(\boldsymbol{\alpha})) \end{aligned} \tag{5.6}$$

where $\mathbf{A} = \int_a^b \boldsymbol{M}^{T''}(t)\boldsymbol{M}''(t)dt$ is an $m \times m$ matrix with (h,l) element $\int_a^b M_h''(t)M_l''(t)dt$, l is the log-likelihood function given in Equation (5.5), and $\lambda \geq 0$ is a smoothing parameter that controls the balance between smoothness of the age-specific relative incidence

function and the fit to the data. The larger the value of λ, the smoother the age effect.

As discussed in Section 2.2, the function $\psi(t)$ is not identifiable without some further constraint as it represents a relative effect. In the present setting, it is convenient to use the constraint $\int_a^b \psi(t)dt = 1$, so that φ in Equation (2.8) is the average baseline incidence over the interval $(a, b]$. In Farrington and Whitaker (2006), the constraint $\psi(a) = 1$ was used, so that φ is the baseline incidence at a. Therefore, to obtain the parameter estimates, the penalised log-likelihood function (5.6) is maximised for fixed λ under the constraint $\int_a^b \psi(t)dt = 1$. The cumulative relative incidence is represented as the integral of a linear combination of cubic M-spline functions of the form $\int_a^t (\sum_{l=1}^m g(\alpha_l) M_l(u))du$. Since the integral of an M-spline is an I-spline, the cumulative incidence is represented by a linear combination of I-splines of the form $\sum_{l=1}^m g(\alpha_l) I_l(t)$. From the definition of I-splines all the I_l's evaluated at $t = b$ are equal to 1. Hence the required constraint can be achieved by constraining the sum of the coefficients of the linear combination of cubic M-spline functions to be 1. That is, $\sum_{l=1}^m g(\alpha_l) = \sum_{l=1}^m \alpha_l^2 = 1$.

5.2.2 Smoothing Parameter Selection

The smoothing parameter λ can be provided by the user or selected using automatic methods. We use a cross-validation method as in Joly *et al.* (1998), in which an approximate cross-validation score is maximised with β set to zero. Let $\boldsymbol{\alpha}$ be the vector of parameters α_l. Denote the cross-validation score $V(\lambda)$,

$$V(\lambda) = \sum_i^N l_i(\hat{\boldsymbol{\alpha}}_{-i}) \tag{5.7}$$

where $\hat{\boldsymbol{\alpha}}_{-i} = \hat{\boldsymbol{\alpha}}_{-i}(\lambda)$ is the maximum penalised log-likelihood estimator of $\boldsymbol{\alpha}$ (with $\beta = 0$) when individual i is removed, and l_i is the log-likelihood contribution of individual i. Using a first order Taylor approximation around $\hat{\boldsymbol{\alpha}}$, the penalised maximum likelihood

estimate when all observations are included, we get:

$$V(\lambda) = \sum_{i=1}^{N} l_i(\hat{\boldsymbol{\alpha}}_{-i}) \approx \sum_{i=1}^{N} \left(l_i(\hat{\boldsymbol{\alpha}}) + \frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}})[\hat{\boldsymbol{\alpha}}_{-i} - \hat{\boldsymbol{\alpha}}] \right)$$

Following O'Sullivan (1988a) $\hat{\boldsymbol{\alpha}}_{-i}$ can approximated as: $\hat{\boldsymbol{\alpha}}_{-i} \approx \hat{\boldsymbol{\alpha}} - [\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{\mathbf{d}}_{-i}$, where $\hat{H} = \frac{\partial^2 l}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}(\hat{\boldsymbol{\alpha}})$ is the log-likelihood part of the Hessian of the penalised log-likelihood evaluated at $\hat{\boldsymbol{\alpha}}$, $\hat{\mathbf{d}}_{-i}^T = -\hat{\mathbf{d}}_i^T = -(\frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}))$ is a score vector when individual i is removed and $2\lambda\mathbf{S}$ is the penalised part of the Hessian, that is $2\lambda\mathbf{S} = \frac{\partial^2(\lambda((g(\boldsymbol{\alpha}))^T \mathbf{A} g(\boldsymbol{\alpha})))}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}$.

Therefore $V(\lambda)$ is approximated by $\bar{V}(\lambda)$, where

$$\begin{aligned}
\bar{V}(\lambda) &= \sum_{i=1}^{N} \left(l_i(\hat{\boldsymbol{\alpha}}) + \frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}})[\hat{\boldsymbol{\alpha}} - [\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{\mathbf{d}}_{-i} - \hat{\boldsymbol{\alpha}}] \right) \\
&= l(\hat{\boldsymbol{\alpha}}) + \sum_{i=1}^{N} -\hat{\mathbf{d}}_i^T [[\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{\mathbf{d}}_{-i}] \\
&= l(\hat{\boldsymbol{\alpha}}) + \sum_{i=1}^{N} \hat{\mathbf{d}}_{-i}^T [\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{\mathbf{d}}_{-i} \\
&= l(\hat{\boldsymbol{\alpha}}) + \sum_{i=1}^{N} \mathrm{tr}(\hat{\mathbf{d}}_{-i}^T [\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{\mathbf{d}}_{-i}) \\
&= l(\hat{\boldsymbol{\alpha}}) + \sum_{i=1}^{N} \mathrm{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{\mathbf{d}}_{-i}\hat{\mathbf{d}}_{-i}^T) \\
&= l(\hat{\boldsymbol{\alpha}}) + \mathrm{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1} \sum_{i=1}^{N} \hat{\mathbf{d}}_{-i}\hat{\mathbf{d}}_{-i}^T) \\
&= l(\hat{\boldsymbol{\alpha}}) + \mathrm{tr}\left([\hat{H} - 2\lambda\mathbf{S}]^{-1} \sum_{i=1}^{N} \left(-\frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}) \right)^T \left(-\frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}) \right) \right) \\
&= l(\hat{\boldsymbol{\alpha}}) + \mathrm{tr}\left([\hat{H} - 2\lambda\mathbf{S}]^{-1} \sum_{i=1}^{N} \left(\frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}) \right)^T \left(\frac{\partial l_i}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}) \right) \right)
\end{aligned}$$

Under regularity conditions $E\left(\frac{\partial l_i}{\partial \alpha_l}\frac{\partial l_i}{\partial \alpha_m}\right) = -E\left(\frac{\partial^2 l_i}{\partial \alpha_l \partial \alpha_m}\right)$, hence

$$V(\lambda) \approx \bar{V}(\lambda) = l(\hat{\boldsymbol{\alpha}}) - \mathrm{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{H}). \tag{5.8}$$

The matrix $\mathbf{S}$ depends on $g(\alpha_l)$. If $g(\alpha_l) = \alpha_l$ then $\mathbf{S} = \mathbf{A}$, where $\mathbf{A}$ is as defined in Section 5.2.1. But here we take $g(\alpha_l) = \alpha_l^2$, therefore denoting point wise product of

matrices by o

$$\mathbf{S} = 4\left(\mathbf{A}\mathrm{o}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\right) + 2(\mathrm{diag}(\mathbf{A}\boldsymbol{\alpha}^2)).$$

This can be shown as follows:

Let $p(\boldsymbol{\alpha}) = (g(\boldsymbol{\alpha}))^T \mathbf{A} g(\boldsymbol{\alpha})$ and $\boldsymbol{\theta} = g(\boldsymbol{\alpha})$. If $g(\alpha_l) = \alpha_l^2$ then, since $\mathbf{A}$ is symmetric, we have

$$\begin{aligned}\mathbf{S} = \frac{1}{2}\frac{\partial^2 p(\alpha)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} &= \left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\alpha}}\right)^T \mathbf{A}\left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\alpha}}\right) + (\boldsymbol{\theta}^T\mathbf{A}\otimes\mathbf{I})\frac{\partial}{\partial\boldsymbol{\alpha}^T}\left[\mathrm{vec}\left(\frac{\partial\boldsymbol{\theta}^T}{\partial\boldsymbol{\alpha}}\right)\right] \\ &= 4\left(\mathbf{A}\mathrm{o}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\right) + 2(\mathrm{diag}(\mathbf{A}\boldsymbol{\alpha}^2)). \qquad (5.9)\end{aligned}$$

The quantity $\mathrm{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{H})$ can be interpreted as a model degrees of freedom. We checked the validity of the approximation in simulation studies (see Section 5.3). The penalised log-likelihood function (5.6) is maximised, with $\beta = 0$, for a grid of λ values, and the value of λ that maximises the approximate cross-validation score (5.8), is used in a final optimisation step with the full model to obtain the relative incidences related to exposure.

5.2.3 Fitting the Spline-Based SCCS Model

The information needed to fit the spline-based SCCS model is the same as for the standard SCCS. Instead of selecting age groups, a suitable (large) number of knots is chosen. A large number of knots is chosen deliberately to over-fit, and we use a penalised log-likelihood to control the balance between the smoothness and roughness of the fitted curve. Usually between eight and 15 knots are sufficient (Rondeau and Gonzalez, 2005; Joly *et al.*, 1998). The knots will include the values a and b, namely, the minimum age at the start of all observation periods and maximum age at the end of all observation periods. The number of knots will depend to some extent on the degree of age variation - the more variation, the more knots should be used. The knots can be equidistant or

chosen based on quantiles of the event times. These choices usually have little impact on results.

In a first step, λ is chosen using the approximate cross-validation method ignoring the exposure effect. Then, the parameters are estimated by maximising the penalised log-likelihood function (5.5) with the chosen value of λ, and under the constraint that the coefficients of the age-specific relative incidence function sum to one. The inverse of the Hessian of the penalised log-likelihood is used as a variance estimator of the parameters (Rondeau and Gonzalez, 2005). These standard errors are Bayesian related standard errors proposed by O'Sullivan (1988a), considering the penalty term in the penalised log-likelihood function (5.6) as a prior. For more detail, see Section 6.1.1 of Chapter 6.

Multiple risk periods can readily be incorporated. In addition to incorporating an indicator for the new exposure status in the numerator of the log-likelihood function (5.5) we add $(\exp(\rho)-1)(\sum_{l=1}^{m}\alpha_l^2 I_l(e_i)-\sum_{l=1}^{m}\alpha_l^2 I_l(s_i))$ in its denominator, where $\exp(\rho)$ is the relative incidence of the new exposure and s_i and e_i are the ages at the start and end of the risk period associated with the new exposure for individual i, respectively. The log-likelihood function with two risk periods is given as

$$l=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\left(\frac{(\sum_{l=1}^{m}\alpha_l^2 M_l(t_{ij}))\exp\{x_i(t_{ij})\beta\}\exp\{y_i(t_{ij})\rho\}}{I(b_i)-I(a_i)+(\exp(\beta)-1)(I(d_i)-I(c_i))+(\exp(\rho)-1)(I(e_i)-I(s_i))}\right), \tag{5.10}$$

where $y_i(t_{ij})$ is 1 if t_{ij} is within the risk period $(s_i, e_i]$ and 0 otherwise. $I(e_i)$ and $I(s_i)$ are linear combinations of I-splines given by $I(s_i)=\sum_{l=1}^{m}\alpha_l^2 I_l(s_i)$ and $I(e_i)=\sum_{l=1}^{m}\alpha_l^2 I_l(e_i)$ respectively. Further exposures can be added in the same way; some care is required in handling overlapping risk periods.

The new spline-based SCCS method has been implemented in R 2.15.1 R Development Core Team (2012), and the optimisation of the constrained penalised log-likelihood

maximisation is done using R function *auglag* from package *alabama.*

5.3 Simulation Study

We conducted a simulation study to investigate the performance of the new method and to compare it with the piecewise constant and semi-parametric versions of the SCCS method. We also made a comparison of the new method with SCCS with no age effect included in the model. The simulation study evaluated how well the new method estimates the age-specific relative incidence function and exposure-related relative incidence.

5.3.1 Design of the Simulation Study

The parameters which need to be specified in the simulation are: the observation period (age at start of observation and age at end of observation), the distribution of age at exposure, c_i, the length of the post exposure period, the exposure-related relative incidences, the age-specific relative incidence function, and sample size.

The observation periods for all cases were taken to be from age of 0 to 500 days, that is $a_i = 0$ and $b_i = 500$ for all i. We assumed only one exposure period $(c_i, d_i]$ of length 50 days, so $d_i = c_i + 50$, where c_i is age at first exposure of individual i.

The exposure variable $x_i(t)$ takes the value 1 in $(c_i,\ c_i + 50]$ and 0 elsewhere. Three different scenarios for the distribution of ages at start of exposure (c_i) were considered: (a) exponentially decreasing (c_i sampled within [0,500] from an exponential density with rate 0.02), uniformly distributed (c_i sampled from $U[0, 500]$) and exponentially increasing ($500 - c_i$ sampled within $[0, 500]$ from an exponential density with rate 0.02).

The baseline incidence at age t was defined to be $\lambda_0(t) \propto 5\exp(\delta t) + 2$, and three scenarios were investigated: exponentially decreasing (with $\delta = -0.003$), constant ($\delta = 0$)

and exponentially increasing ($\delta = 0.003$). That is, at each day of age in $(0, 500]$ age-specific relative incidences were generated.

True exposure-related relative incidences of 1, 2, and 5 were investigated, with sample sizes 50, 100 and 200. In the simulations just one event was simulated per individual; this involved no loss of generality as multiple events within an individual are treated as independent in SCCS.

We then generated age at event for each individual conditional on the exposure status and number of events an individual experiences, one event per individual in this case, from multinomial distributions with daily categories. 10,000 samples were generated for all combinations of the scenarios.

The new spline-based, semi-parametric and standard SCCS methods were fitted to each of the generated samples. Nine age groups, with cut-points at 30, 92, 155, 218, 281, 344, 407 and 470 days, were used to fit the piecewise constant SCCS model. We also fitted an SCCS without age effect to quantify the bias in the exposure effect when age is ignored. In fitting the spline-based SCCS we used M-splines of order 4 and the number of interior knots was chosen to be 10 hence the number of basis functions was $m = 14$. The values of a and b respectively are 0 and 500 days.

We evaluated the performance of the new method in terms of its fit to the true age-specific relative incidence function and in terms of reflecting the true exposure-related relative incidence. For each of the 10,000 samples the mean of the estimated integrated squared errors (MISE) and their standard deviations were calculated. The MISE values computed were for the cumulative age-specific relative incidence function to make it comparable with the estimates obtained from the semi-parametric SCCS method. The

integrated squared errors (ISE) for each sample are defined as:

$$\int_0^{500} (\Psi(t) - \hat{\Psi}(t))^2 dt,$$

where $\Psi(t)$ is true age-specific cumulative relative incidence and $\hat{\Psi}(t)$ is estimated cumulative relative incidence. After fitting the models for each sample we estimated the cumulative relative incidence at each day of age from 0 - 500 and approximated the ISE values as:

$$\sum_{t=0}^{500} (\Psi(t) - \hat{\Psi}(t))^2.$$

We then find the MISE values by taking the mean of the ISE values. To investigate the performance of the new method in estimating the exposure-related relative incidence we computed the mean, median and standard deviation of the 10,000 log relative incidence estimates ($\hat{\beta}$), the coverage probability of the 95% confidence intervals, empirical standard errors and average model-based standard errors of $\hat{\beta}$. We use the median as well as the mean because there is a non zero probability that all events will occur in the risk period only or in the control period only, so that in finite samples the expectation of $\hat{\beta}$, and hence the bias = expectation of $\hat{\beta}$ - true relative incidence (β), is undefined. (All quantities involving expectations should therefore be regarded as having been trimmed, by removing samples resulting in unbounded estimates.)

5.3.2 Results

In this section, we present results of simulations conducted to evaluate and compare the performance of the spline-based SCCS in terms of estimating both the age-specific relative incidence function and the relative incidence related to exposure.

Performance in Estimating the Age Effect

Tables 5.1 and 5.2 show results from the simulation in evaluating the performance of the new method relative to the semi-parametric SCCS model in estimating the cumulative age-specific relative incidence function. MISE values and standard deviations of ISE values are presented.

Table 5.1: *Mean integrated squared Error (MISE) and Standard Deviation (SD) for estimating the cumulative age-specific relative incidence using spline-based and semi-parametric SCCS: simulations based on different scenarios of age at exposure (AE), age-specific relative incidence (ASRI) and exposure relative incidence (RI).*

# of cases	Spline-based MISE(SD)	Semi-parametric MISE(SD)	Spline-based MISE(SD)	Semi-parametric MISE(SD)
	ASRI increasing & AE decreasing, RI=5		ASRI & AE decreasing, RI=2	
50	2.004(2.084)	2.208(2.103)	1.911(2.010)	2.102(2.027)
100	0.977(1.012)	1.055(1.012)	0.950(0.962)	1.029(0.965)
200	0.513(0.531)	0.548(0.531)	0.468(0.491)	0.502(0.490)
	ASRI increasing & AE increasing, RI=5		ASRI increasing & AE uniform, RI=1	
50	1.771(2.161)	2.016(2.174)	1.288(1.322)	1.442(1.312)
100	0.910(1.052)	1.006(1.054)	0.656(0.679)	0.724(0.679)
200	0.444(0.538)	0.484(0.541)	0.326(0.344)	0.354(0.343)

We can see from Table 5.1 that the performance of the new method in approximating the true age-specific relative incidence is similar to though slightly better than the semi-parametric method as the MISE values are slightly lower for the new method. The reduction in the MISE is of the order of 7%, representing a slight gain in efficiency. The results presented in Table 5.1 show how the MISE values vary with a change in the number of cases for different scenarios of age at exposure, relative incidence and age-specific relative incidence function. The MISE values in Table 5.1 decrease with an

increase in the number of cases used in the simulations.

Table 5.2: *Mean integrated squared Error (MISE) and Standard Deviation (SD) for estimating age-specific relative incidence using spline-based and semi-parametric SCCS: simulations based on two scenarios of age at exposure (AE), age-specific relative incidence (ASRI) and exposure relative incidence (RI) for 50 and 100 cases.*

	50 cases		100 cases	
True RI	Spline-based MISE(SD)	semi-parametric MISE(SD)	Spline-based MISE(SD)	semi-parametric MISE(SD)
Constant ASRI and increasing AE				
1	1.469(1.738)	1.615(1.744)	0.706(0.847)	0.771(0.849)
2	1.576(1.719)	1.736(1.728)	0.764(0.908)	0.835(0.908)
5	1.697(2.008)	1.896(2.003)	0.888(1.136)	0.974(1.137)
Increasing ASRI and Uniform AE				
1	1.312(1.517)	1.458(1.514)	0.668(0.723)	0.736(0.723)
2	1.252(1.458)	1.408(1.456)	0.654(0.703)	0.722(0.699)
5	1.340(1.474)	1.507(1.486)	0.683(0.767)	0.755(0.768)
Decreasing ASRI and Increasing AE				
1	1.465(1.663)	1.618(1.619)	0.744(0.847)	0.807(0.828)
2	1.575(1.845)	1.724(1.800)	0.759(0.795)	0.822(0.787)
5	1.589(1.826)	1.772(1.831)	0.846(1.017)	0.924(1.001)

Results for scenarios where the age-specific relative incidence function is constant, exponentially increasing and decreasing with uniformly distributed and exponentially increasing age at exposure are presented in Table 5.2. Results relating to the comparison of spline-based and semi-parametric methods in this Table are similar to the ones presented in Table 5.1. In general, the Mean Integrated Squared Error increases with an increase in the true exposure-related relative incidence value. For both the spline-based and semi-parametric methods, the efficiency in estimating the age-specific relative incidence function increases with an increase in sample size, because the MISE values decrease

as the number of cases increase. The spline-based method has a slightly higher efficiency in estimating an exponentially increasing or decreasing age effect than a constant function.

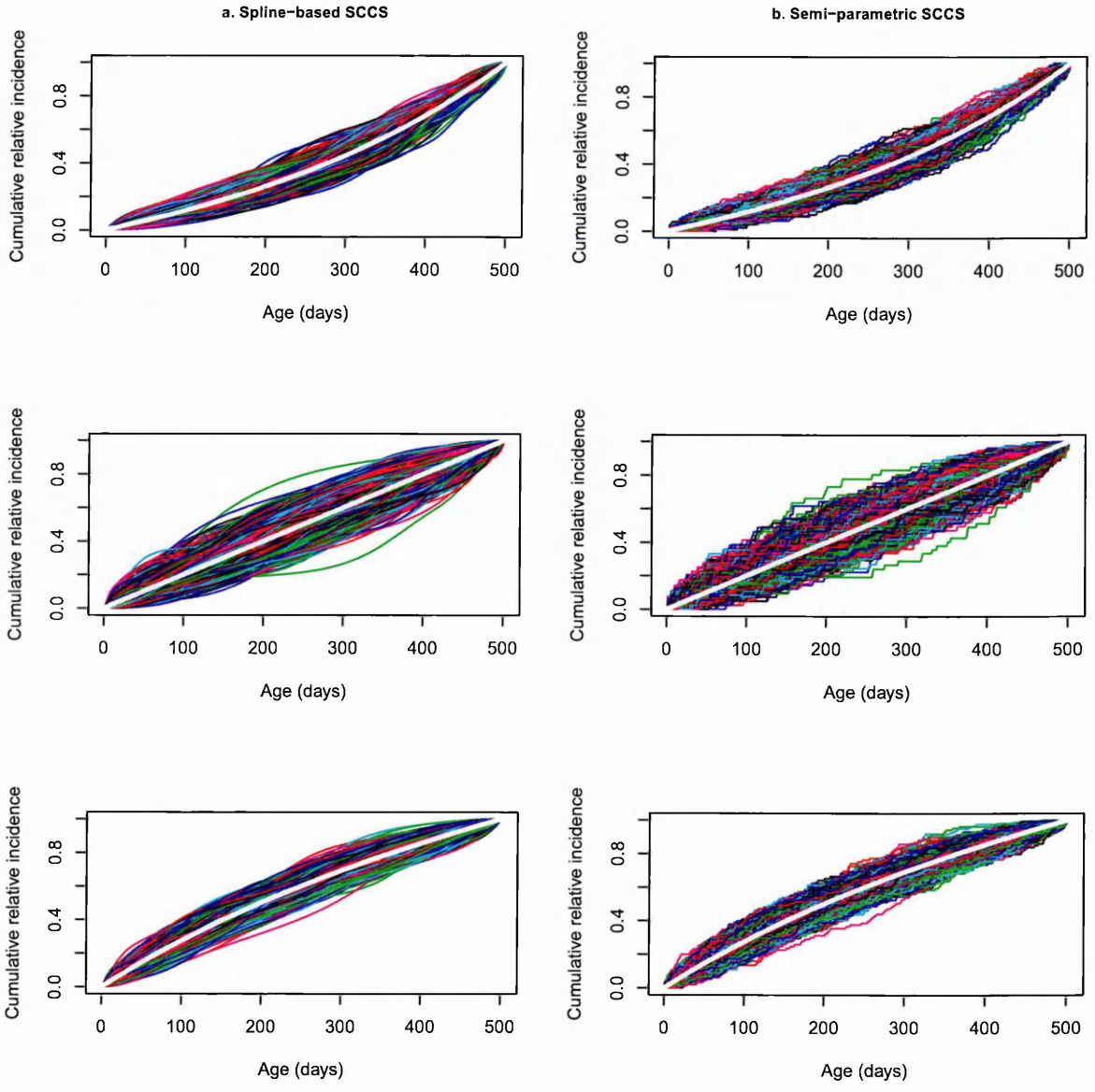

Figure 5.3: *Estimated cumulative age-specific relative incidence curves of the first 1,000 simulated data sets, Panel a represent results from spline-based method and Panel b results from semi-parametric SCCS. In the top panels the true curve is exponentially increasing, in the middle panels a constant and in the bottom panels an exponentially decreasing function*

Figure 5.3 shows estimated age-specific cumulative relative incidence curves from the

spline-based and semi-parametric methods fitted to the first 1,000 simulated data sets. The white bold lines represent the true curves. The top panels show results when the true age-related relative incidence function is exponentially increasing, the middle panels when it is a constant function and the bottom panels are when the true curve is an exponentially decreasing. Results in panel *a* of Figure 5.3 are from the spline-based method and results in panel *b* are from the semi-parametric SCCS method.

From Figure 5.3 both the spline-based and semi-parametric SCCS methods seem to approximate the true age-specific relative incidence functions well. The variabilities in estimating the constant function seem to be more than the exponentially increasing and decreasing age-specific relative incidence functions.

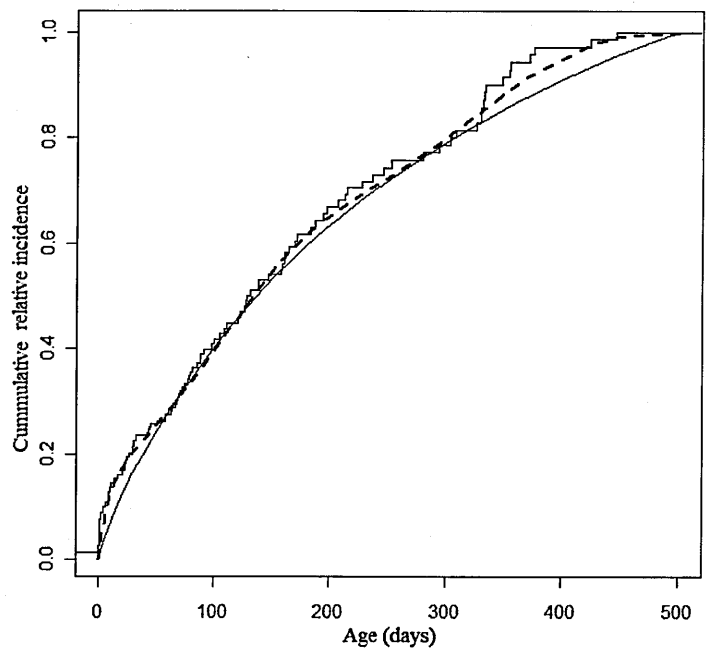

Figure 5.4: *Cumulative age-specific relative incidence curves for a single simulated sample: True curve (bold line), a curve estimated using spline-based SCCS from a single simulated data set (dashed line) and the step function estimated using the semi-parametric model from the same single simulated data set.*

Figure 5.4 shows the cumulative age-specific relative incidence from the spline-based SCCS, the semi-parametric SCCS and the true cumulative incidence for a single sample of simulated data of 200 cases. From the plot it seems that the two estimation methods give similar results for this data set.

For this same sample of simulated data, we evaluated the approximate cross-validation score given in Equation (5.8) at a smoothing parameter (λ) value of 100,000. We also calculated the exact cross-validation score by leaving out one case at a time and fitting the spline-based SCCS iteratively. The two values were close to each other, 1,182.798 and 1,183.045 respectively.

Performance in Estimating the Exposure Effect

The mean and median of the log exposure-related relative incidence $\log(\text{RI}) = \beta$, empirical standard errors and average model-based standard errors of $\hat{\beta}$, for a scenario in which both the age-specific relative incidence function and age at exposure decrease exponentially, are shown in Table 5.3. The coverage probabilities of the 95% confidence intervals for β are also presented. Table 5.3 shows that the bias in estimating the exposure-related relative incidence using the new method is small with these sample sizes. It is similar to or smaller than the bias for the semi-parametric method, and smaller than the bias of the parametric method. The empirical and average model-based standard errors for the spline method lie between those achieved by the parametric method (which are lower) and the semi-parametric method (higher). The spline-based confidence intervals tend to be slightly conservative, though not as badly so as those for the parametric method. As expected, the SCCS without age effect produces very biased results and has a generally poor performance, underlining the importance of age adjustment in SCCS models.

Table 5.3: *Simulation results from a scenario where age at exposure and age-specific relative incidence decrease exponentially. Mean, Median, Empirical standard errors (ESE), Average model based standard error (AMSE) and 95% coverage probability (P95) for the log relative incidence, log(RI) = β, are presented.*

Spline-based SCCS										
log(RI)	50 cases					100 cases				
	Median	Mean	ESE	AMSE	P95	Median	Mean	ESE	AMSE	P95
1.609	1.657	1.690	0.486	0.429	94.107	1.624	1.637	0.287	0.269	94.614
0.693	0.697	0.707	0.394	0.371	94.482	0.697	0.701	0.274	0.262	95.358
0.000	-0.005	-0.009	0.442	0.408	94.343	-0.005	-0.004	0.448	0.412	94.098
semi-parametric SCCS										
1.609	1.657	1.700	0.512	0.484	95.600	1.630	1.646	0.291	0.281	94.600
0.693	0.709	0.717	0.404	0.398	95.300	0.704	0.706	0.276	0.270	94.700
0.000	0.004	-0.005	0.453	0.440	95.080	-0.004	-0.004	0.460	0.447	95.410
Parametric SCCS										
1.609	1.452	1.478	0.438	0.422	92.520	1.520	1.525	0.273	0.264	92.450
0.693	0.601	0.604	0.400	0.388	94.230	0.605	0.610	0.256	0.255	94.080
0.000	-0.132	-0.144	0.414	0.399	93.640	-0.105	-0.111	0.288	0.285	94.020
SCCS without age effect										
1.609	2.140	2.165	0.305	0.309	56.020	2.125	2.121	0.217	0.215	33.540
0.693	1.181	1.185	0.288	0.288	59.480	1.122	1.124	0.202	0.203	42.970
0.000	0.539	0.534	0.320	0.314	56.280	0.539	0.545	0.314	0.313	54.930

Table 5.4 presents results for a scenario where ages at exposure are uniformly distributed and age-specific relative incidence function increases exponentially. In this scenario there is less confounding effect of age on the exposure-related relative incidence as the chance to get exposed does not depend on age. As a consequence the SCCS method with age excluded in the model gave better results as compared to the previous scenario where age at exposure and age-specific relative incidence decrease exponentially. It has

less bias and the coverage probability is closer to 95%. However ignoring age still has lower performance than the semi-parametric and spline-based SCCS methods. The spline method shows similar performance to the previous scenario, that is as good as or better than the semi-parametric and the standard methods.

Table 5.4: *Simulation results from a scenario where age at exposure is uniformly distributed and age-specific relative incidence function increases with age exponentially. Mean, Median, Empirical standard errors (ESE), Average model based standard error (AMSE) and 95% coverage probability (P95) for the log relative incidence, log(RI) = β, are presented.*

Spline-based SCCS										
log(RI)	50 cases					100 cases				
	Median	Mean	ESE	AMSE	P95	Median	Mean	ESE	AMSE	P95
1.609	1.626	1.611	0.336	0.325	94.660	1.625	1.622	0.235	0.231	94.852
0.693	0.664	0.648	0.423	0.402	95.300	0.689	0.674	0.282	0.281	96.400
0.000	-0.024	-0.087	0.528	0.515	96.690	-0.045	-0.078	0.376	0.365	96.400
semi-parametric SCCS										
1.609	1.648	1.653	0.352	0.337	94.114	1.641	1.638	0.240	0.234	94.604
0.693	0.688	0.663	0.430	0.408	94.800	0.695	0.683	0.285	0.283	96.100
0.000	-0.019	-0.081	0.532	0.519	96.790	-0.042	-0.075	0.377	0.367	96.400
Parametric SCCS										
1.609	1.597	1.589	0.340	0.326	94.600	1.598	1.595	0.235	0.232	95.021
0.693	0.659	0.638	0.419	0.403	96.100	0.667	0.659	0.283	0.282	96.200
0.000	-0.031	-0.087	0.536	0.517	96.891	-0.046	-0.077	0.375	0.365	96.300
SCCS without age effect										
1.609	1.549	1.533	0.314	0.311	93.871	1.521	1.520	0.220	0.221	94.379
0.693	0.628	0.647	0.404	0.393	96.400	0.671	0.672	0.275	0.278	96.200
0.000	0.140	0.053	0.521	0.508	94.383	-0.046	-0.082	0.370	0.362	97.100

While the spline-based, the semi-parametric and the standard methods show similar performance, when the effect of age is kept constant and age at exposure increases expo-

nentially, the SCCS without age effect performed well. And this is because keeping the age-specific relative incidence function to be a constant function means that the effect of age is cancelled out in the SCCS log-likelihood function. That is the true model is an SCCS model with no age effect. Results for this scenario are presented in Table 5.5.

Table 5.5: *Simulation results from a scenario where age at exposure increases exponentially and age-specific relative incidence function is constant. Mean, Median, Empirical standard errors (ESE), Average model based standard error (AMSE) and 95% coverage probability (P95) for the log relative incidence, log(RI) = β, are presented.*

Spline-based SCCS										
log(RI)	50 cases					100 cases				
	Median	Mean	ESE	AMSE	P95	Median	Mean	ESE	AMSE	P95
1.609	1.621	1.633	0.471	0.483	97.080	1.640	1.645	0.368	0.360	94.700
0.693	0.667	0.643	0.666	0.630	95.996	0.650	0.643	0.440	0.436	96.200
0.000	-0.047	-0.062	0.729	0.779	97.092	-0.004	-0.063	0.571	0.538	97.189
semi-parametric SCCS										
1.609	1.635	1.653	0.482	0.492	96.982	1.646	1.653	0.372	0.363	93.894
0.693	0.687	0.659	0.677	0.639	95.796	0.646	0.655	0.444	0.439	95.800
0.000	-0.015	-0.041	0.743	0.787	97.508	0.005	-0.045	0.571	0.541	96.586
Parametric SCCS										
1.609	1.563	1.575	0.476	0.477	96.797	1.562	1.574	0.361	0.354	94.700
0.693	0.640	0.618	0.654	0.624	96.697	0.617	0.617	0.440	0.429	96.300
0.000	0.016	-0.035	0.719	0.771	97.399	-0.001	-0.051	0.560	0.531	96.790
SCCS without age effect										
1.609	1.596	1.549	0.350	0.353	95.000	1.557	1.564	0.253	0.249	95.700
0.693	0.556	0.563	0.531	0.498	96.296	0.629	0.611	0.356	0.349	96.100
0.000	-0.004	-0.099	0.599	0.657	96.982	0.051	-0.080	0.506	0.467	97.392

5.4 Analysis of Febrile Convulsion Data

We apply the new spline-based SCCS method to data on febrile convulsions and paediatric vaccines collected in England and Wales in the period of 1991–1994. SCCS with piecewise constant age-specific relative incidence function was also applied for comparison purposes.

The aim of the analysis is to investigate the association between febrile convulsions and diphtheria/tetanus/pertussis (DTP), Haemophilus influenza type B (Hib) and measles/ mumps/ rubella (MMR) vaccines. Febrile convulsions or seizures are a relatively common childhood condition, referring to a child having a seizure (fit) when they have a high temperature. It occurs when the electrical impulses, used to communicate brain cells (neurons), become disrupted. This can cause the brain and the body to behave abnormally (NHS, 2013).

DTP vaccine is given to prevent three diseases: diphtheria, tetanus, pertussis (whooping cough), and the Hib vaccine is a vaccine developed to prevent invasive disease caused by Haemophilus influenzae type b bacteria. The disease is a bacterial infection that can cause a number of serious illnesses such as pneumonia or meningitis, especially in young children (NHS, 2013). The Hib vaccination is offered to children at two, three and four months of age, which we denote here as Hib1 Hib2 and Hib3 respectively. It is usually given along with DTP and polio vaccines. Another type of Hib vaccine is also given if the first three doses are missed, denoted here as Hibonly. MMR is a combined vaccine that protects against measles, mumps and rubella (German measles). Measles, mumps and rubella are very common, highly infectious conditions that can have serious, potentially fatal, complications, including meningitis, swelling of the brain (encephalitis) and

deafness (NHS, 2013). The first MMR vaccine is given usually within 12 to 13 months of age and a second dose of the vaccine is given between ages of three and five. Here we only study the effect of the first MMR dose in causing febrile convulsion.

The data set includes 2,389 children aged 29–730 days in the period 1991 to 1994, who had 3,826 febrile convulsion events. Of the 2,389 cases, 2,021 cases had an MMR vaccine record. DTP vaccine was given in three doses, DTP1, DTP2 and DTP3. The number of cases vaccinated with DTP1, DTP2 and DTP3 were 1,624, 1,684 and 1,726 respectively. And the numbers of Hib vaccinated children were 1,706, 1,636, 1,552 and 880 for Hib1, Hib2, Hib3 and Hibonly respectively.

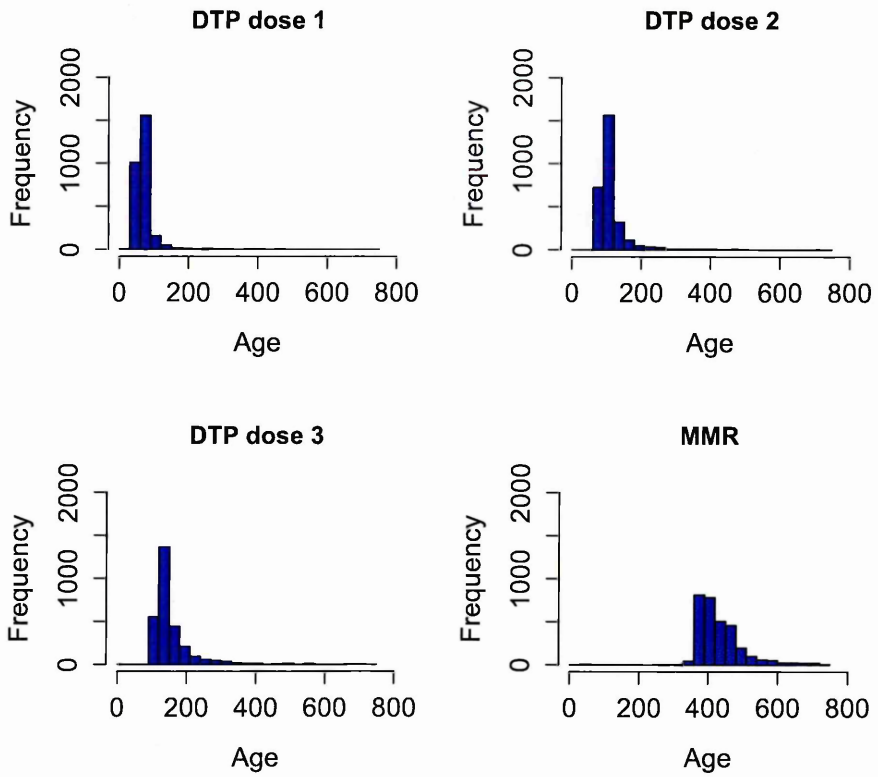

Figure 5.5: *The distribution of age at DTP and MMR Vaccines. DTP was taken in three doses*

The average ages at which DTP1, DTP2, DTP3 and MMR vaccines were taken are

74.3, 119.1, 167.7 and 437.0 days respectively. The mean of ages at Hib1, Hib2, Hib3 and Hibonly are 136.1, 177.0, 221.9, and 520.9 days respectively. The distributions of ages at exposure to the DTP and MMR vaccines are presented in Figure 5.5. Figure 5.6 presents the distributions of ages at Hib vaccines. The figures show that the chances of being exposed to the vaccines depend strongly on age hence age might have a confounding effect if it is related to the rate of baseline incidence too. Therefore appropriate modelling of age effect is desirable.

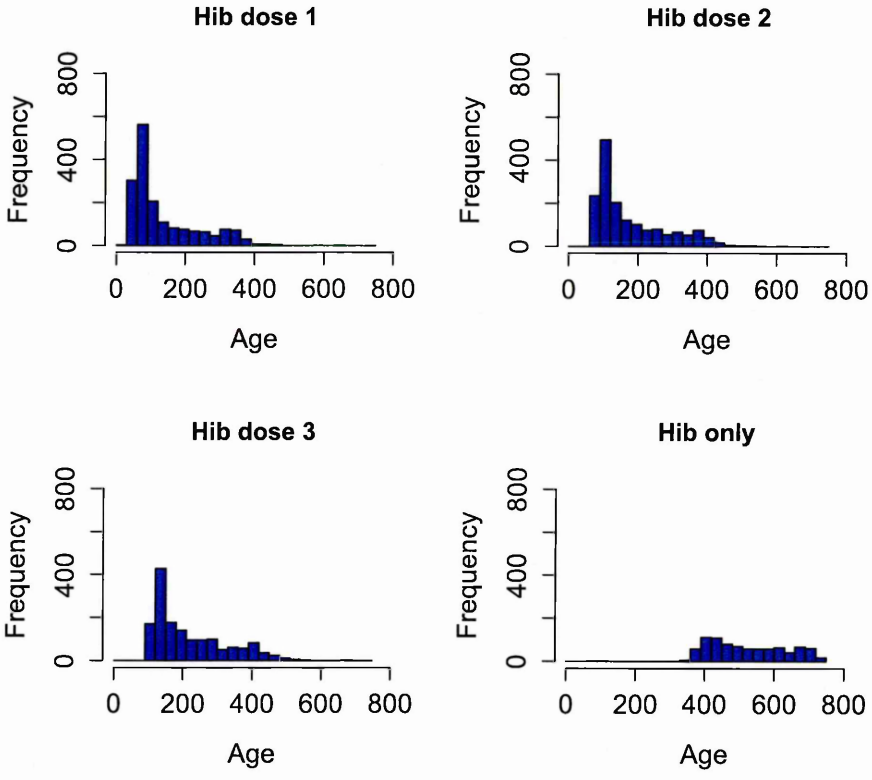

Figure 5.6: *The distribution of age at exposure to Hib vaccine. Hib was given in three doses and one dose Hibonly if the first three are missed*

Febrile convulsions are relatively rare but potentially recurrent events, most cases experiencing a single convulsion over the two years. Furthermore, febrile convulsions are not contra-indications to vaccination, and carry a very low mortality risk. Thus, there is

no reason to doubt that conditions (1) to (3) of Section 2.1.1 in Chapter 2 are valid for these data.

The large number of cases precludes us from using the semi-parametric method because the method does not work for large number of cases as seen in the simulation study conducted in Section 2.3 of Chapter 2. We estimated relative incidences (RI) for febrile convulsion in risk periods following these vaccines compared to control periods, using both the spline-based and standard SCCS methods.

In a first analysis, we considered exposures to DTP and MMR vaccines. We estimated the relative incidences associated with exposure to the three doses of DTP (DTP1, DTP2, DTP3) using risk periods of 0–7 days for all the doses and MMR vaccine with a risk period of 6–11 days post vaccination. Exposure status was represented by four time-varying indicator variables, taking the value 1 in the relevant risk period and 0 in the control periods.

Overlapping risk periods were coded to the latest vaccine (Whitaker *et al.* (2006)). For the standard method, age was divided into 23 equal intervals of 1 month, apart from the first age group which had 32 days and the last 40 days. The values of a and b respectively are the minimum age at start of observations (29 days) and maximum age at end of observation of the cases (730 days). In the spline-based SCCS analysis the age-specific relative incidence function was approximated by a linear combination of cubic M-splines with 14 knots (12 interior plus a and b). Therefore, the number of M-spline functions was $m = 16$. The internal knots were at roughly equal intervals. We maximised the penalised log-likelihood function (5.6), excluding the covariates that represent DTP and MMR vaccines, at different values of the smoothing parameter λ. The value of λ at which the approximated cross-validation score is maximum was found to be 1.07×10^9. We then

maximised the penalised log-likelihood function (5.6), fixing λ at 1.07×10^9 to obtain the relative incidences of febrile convulsion related to DTP1, DTP2, DTP3 and MMR vaccines. Results of these analysis are presented in Table 5.6

Table 5.6: *Relative incidence (RI) and 95% Confidence intervals (CI) for febrile convulsion due to exposure to three doses of DTP and MMR vaccines. Three parameter estimates for DTP and one for MMR for the risk period of 6 -11 days after vaccination*

		Spline-based SCCS		Standard SCCS	
Vaccine	Risk period	RI	95% CI	RI	95% CI
DTP1	0-7	1.068	[0.643 , 1.775]	1.26	[0.775 , 2.048]
DTP2	0-7	0.995	[0.584 , 1.696]	0.937	[0.542 , 1.619]
DTP3	0-7	1.413	[0.923 , 2.160]	1.293	[0.830 ,2.014]
MMR	6-11	3.214	[2.656 , 3.885]	3.386	[2.806 , 4.088]

Table 5.6 shows that results from both the spline-based and standard SCCS methods are similar. There was no significantly increased risk of febrile convulsion of the three doses of DTP whereas exposure to MMR vaccine had a significant effect with a relative incidence of 3.214[2.656 , 3.885] and 3.386[2.806 , 4.088] from spline-based and standard SCCS analysis respectively.

Table 5.7: *Relative incidence (RI) and 95% Confidence intervals (CI) for febrile convulsion due to exposure to DTP and MMR vaccines.*

		Spline-based SCCS		Parametric SCCS	
Vaccine	Risk period	RI	95% CI	RI	95% CI
DTP all doses	0-3	1.905	[1.349 , 2.668]	1.420	[0.963 , 2.092]
	4-7	1.391	[0.933 , 2.075]	1.184	[0.774 , 1.812]
	8-14	1.225	[0.899 , 1.670]	0.974	[0.693 , 1.366]
MMR	6-11	3.781	[3.120 , 4.492]	3.451	[2.854 , 4.175]
	15-35	1.241	[1.050 , 1.453]	1.197	[1.013 , 1.414]

The second analysis was performed by considering any DTP vaccine with three different risk periods (0-3, 4-7 and 8-14 days after vaccination), and two risk periods after MMR vaccination (6-11 and 15-35 days). Results are presented in Table 5.7 and Figure 5.7.

The fitted age-specific relative incidence curves obtained from the standard and the spline-based SCCS are presented in the left panel of Figure 5.7. The right panel of Figure 5.7 shows the cumulative age-specific relative incidence curves, where the dashed line is from the spline-based SCCS and the solid line from the parametric SCCS. The model degrees of freedom obtained for the optimum smoothing parameter value was 7.962.

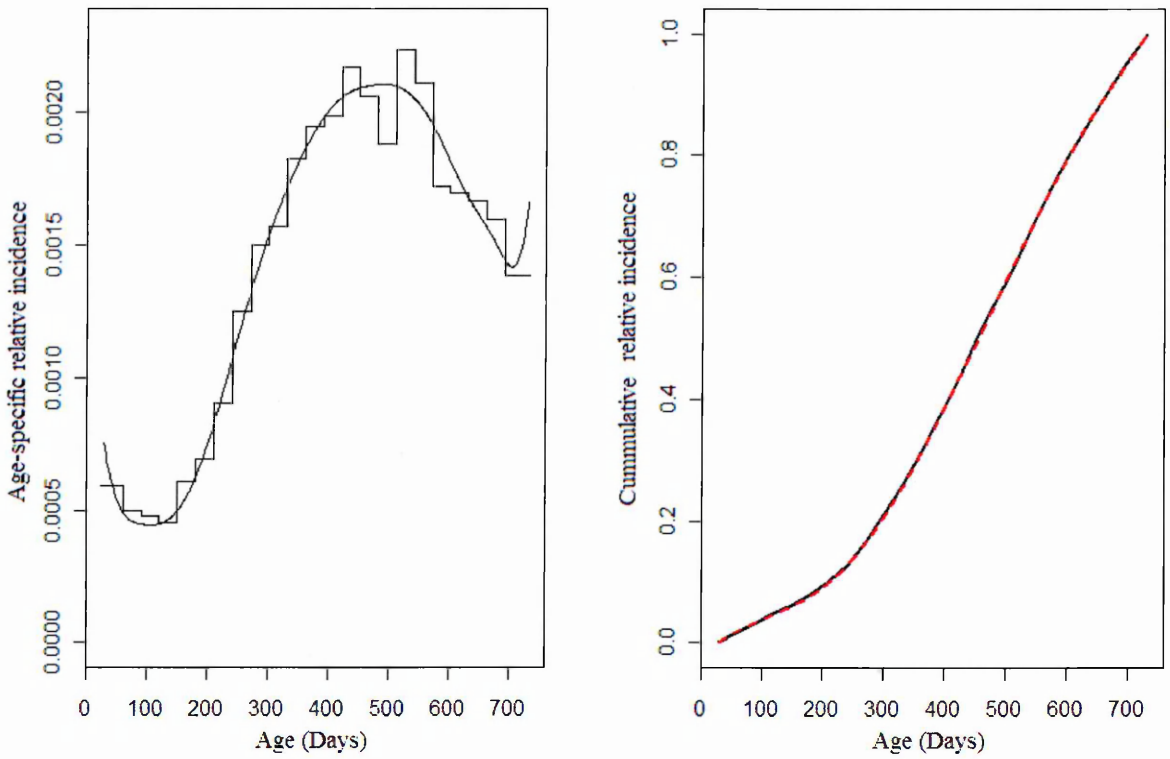

Figure 5.7: *Left: Age-specific relative incidence; step function estimated by parametric SCCS, smooth curve estimated by spline-based SCCS. Right: Cumulative age-specific relative incidence; dashed line estimated by parametric SCCS and solid line estimated by spline-based SCCS.*

Table 5.7 presents exposure-related relative incidence estimates from both methods.

It shows that the two methods gave similar results for MMR with significant associations between febrile convulsion and MMR vaccines in both risk periods, in line with other analyses (Farrington *et al.*, 1995). RI estimates for DTP vaccines in the risk periods 4-7 and 8-14 days were not significantly different from 1 for the two methods. However, there was a difference for the 0-3 days risk period. The RI estimate using the spline-based method was significantly greater than 1 whereas with the standard method it was non-significant. This is due to the very strong age effect in the first year of life, which is inadequately controlled using the standard model with age groups of length one month. Thus, the spline-based method suggests that there may be an association between DTP vaccination and convulsions within 3 days post-vaccination, which was not identified using the piecewise constant model.

The DTP and Hib vaccines are usually given at the same time. In the data set there were large numbers of cases vaccinated with the two vaccines at the same time. The number of cases who took DTP1 and Hib1, DTP2 and Hib2 and DTP3 and Hib3 were 909, 875 and 838 respectively. Since overlapping risk periods were coded to the latest vaccine, we had two final analyses in the investigation of the association between the paediatric vaccines and febrile convulsion. We first assumed that DTP was given before Hib so that the risk period following DTP will be coded by the risk period of Hib if they were both given at the same time. And the second analysis was performed assuming that, for individuals vaccinated with DTP and Hib at the same time, Hib was taken before DTP. In these analyses we used one risk period of 0-3 days post DTP vaccination, one risk period of 0-7 days post Hib, one risk period of 0-7 days post Hibonly vaccine and two risk periods after exposure to MMR vaccine (6-11 and 15-35 days). Results of the analyses are shown in Table 5.8.

Table 5.8: *Relative incidence (RI) and 95% Confidence intervals (CI) for febrile convulsion due to exposure to DTP Hib, Hibonly and MMR vaccines.*

		Spline-based SCCS		Parametric SCCS	
Vaccine	Risk period	RI	95% CI	RI	95% CI
Assuming DTP was taken before Hib					
DTP any dose	0-3	1.551	[0.923 , 2.608]	1.154	[0.633 , 2.106]
Hib any dose	0-7	1.352	[1.026 , 1.783]	1.296	[0.926 , 1.815]
Hibonly	0-7	1.019	[0.549 , 1.893]	1.016	[0.544 1.896]
MMR	6-11	3.794	[3.151 4.537]	3.464	[2.864 , 4.189]
	15-35	1.248	[1.061 1.469]	1.203	[1.018 , 1.420]
Assuming Hib was taken before DTP					
DTP any dose	0-3	1.831	[1.297 , 2.586]	1.431	[0.974 , 2.103]
Hib any dose	0-7	1.131	[0.715 , 1.789]	0.984	[0.555 , 1.745]
Hibonly	0-7	1.045	[0.560 , 1.949]	1.050	[0.563 , 1.960]
MMR	6-11	3.789	[3.157 , 4.546]	3.462	[2.863 , 4.186]
	15-35	1.246	[1.057 , 1.469]	1.202	[1.018 , 1.420]

The relative incidence of exposure to Hibonly vaccine was found to be non-significant in both analyses using both methods (Table 5.8). When DTP was assumed to be given before the Hib vaccine the RI estimate associated with exposure to DTP was found to be non significant whereas exposure to Hib vaccine has an increased risk of febrile convulsion for the spline-based SCCS analysis but not for the standard method. The result was reversed when we assumed Hib vaccine to be given before DTP. Relative incidence of febrile convulsion in the period 0-3 days post DTP vaccine was significantly different from 1 in the spline-based analysis.

To investigate the sensitivity of exposure-related relative incidence estimates with a change in the value of a smoothing parameter, we fitted the spline-based SCCS to the data with different values of the smoothing parameter. We considered any DTP dose vaccine

with three risk periods and MMR vaccines with two risk periods. The relative incidence estimates of exposure to DTP and MMR vaccines with varying smoothing parameter are presented in Figure 5.8.

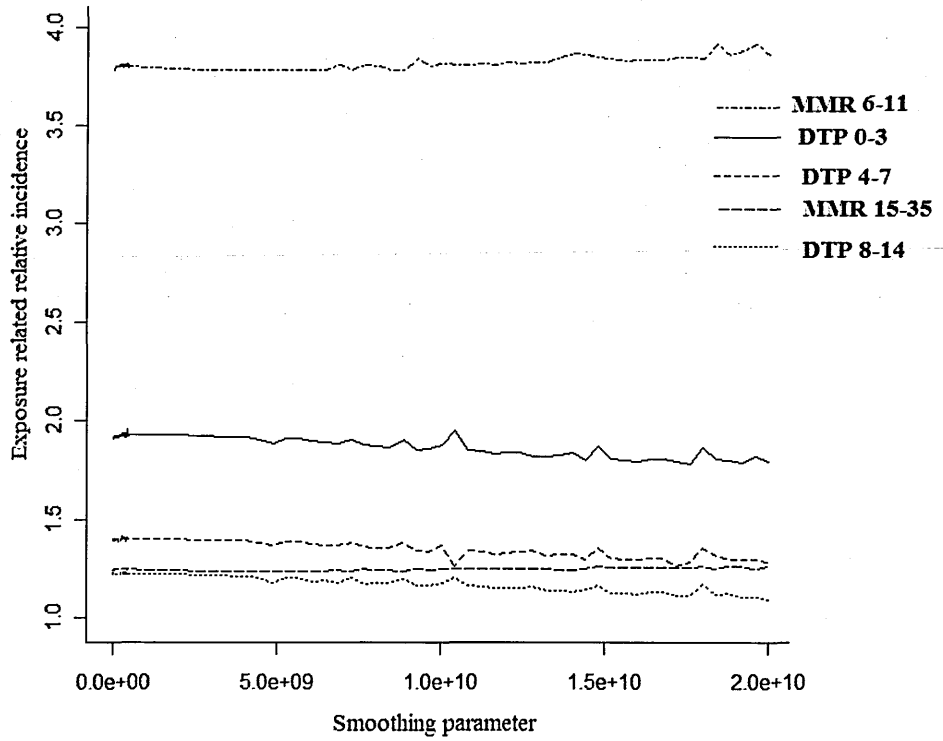

Figure 5.8: *Estimated relative incidence after exposure to MMR and DTP vaccines for specified risk periods (in days: see legend) for different values of the smoothing parameter* λ.

The relative incidence estimates were found to be insensitive to the choice of smoothing parameter. From Figure 5.8, it can be seen that the exposure to vaccine related relative incidence estimates remain similar for varying smoothing parameter values within this range.

5.5 Discussion

Modelling of the age effect in the SCCS model with smooth functions was presented in this Chapter. Using a polynomial function to model the log of the baseline incidence is one way of avoiding the limitations associated with the standard SCCS method. However the use of polynomial functions to represent the log of the age-specific relative incidence function has two limitations: (1) bad fit in one part of a function affects the whole function and (2) for higher order polynomial functions the integral in the denominator of the SCCS log-likelihood function has to be integrated numerically, hence adding to the computational complexity. We suggest to solve problem (2) by directly modelling the age-specific relative incidence function by a polynomial function and constrain the coefficients of the polynomial to be non-negative. This can help to analytically integrate the denominator of the SCCS log-likelihood function.

Lee and Carlin (2014) proposed the use of fractional polynomials to estimate the parameters in the piecewise constant function representing the age effect in the standard SCCS method. However, although the proposed method reduces the number of parameters that need to be estimated compared to the standard and semi-parametric methods, there is a need to pre-define age groups. Moreover, if each day within an observation period is used as an age group it may face the same computational problem as for the semi-parametric method since the data will have to be expanded. However, fractional polynomials can be used to estimate a smooth age effect. We suggest to represent the age-specific relative incidence with fractional polynomials by constraining their coefficients to be non-negative to analytically integrate the denominator of the log-likelihood function. Using this approach the data are not required to be expanded.

Modelling the age effect using a linear combination of cubic M-splines avoids the problem of sensitivity to mis-specification of age groups in the standard (piecewise constant) version of the SCCS method and the computational problem of the semi-parametric SCCS method. The performance of the new method is as good as or better than the semi-parametric and the standard versions of the SCCS method for small and moderate sample sizes. For large samples the semi-parametric SCCS is computationally demanding but the new method works well. For example, for our convulsions data with 3,826 events and 5 risk periods, the spline model took less than two minutes to fit on a standard desktop computer. Part of the problem with fitting the semi-parametric model relates to the use of standard software for fitting Poisson models and maximising conditional likelihoods, which require the data to be expanded. There may be others ways of proceeding that avoid this step. However, when there are N cases, $O(N)$ parameters must be estimated, which is likely to be problematic for large data sets. The spline-based model provides a more economical representation of the age effect in such settings.

Like the standard SCCS methods, the spline-based method requires the three assumptions which were set out in Section 2.1.1 of Chapter 2. As described in Section 5.4, these conditions are likely to be met in our application to febrile convulsions. More generally, if the event of interest is not recurrent, or if recurrences are not independent, then the SCCS model can validly be applied to the first event provided the risk of a first event is low (say less than 10% during a typical observation period (Farrington *et al.*, 2011)). If exposures are not exogenous, then a modified SCCS method can be applied (Farrington *et al.*, 2009). A further version of the method can be used if observation periods are not event-independent (Farrington *et al.*, 2011). On the other hand, experience with the method suggests that results are typically robust to mild departures from the hypothe-

ses. More work is required, however, to characterise the circumstances in which such robustness cannot be taken for granted.

Our aim in the present chapter was to model the age effect non-parametrically, this effect seldom being of primary interest. The exposure effect, on the other hand, is modelled parametrically using step functions, selected on the basis of prior knowledge or hypotheses. In some applications, however, it might be important to estimate the exposure effect more flexibly, notably in exploratory analyses. To this end, we extend the spline-based model to allow flexible modelling of the exposure effect in Chapter 6.

Chapter 6

Flexible Modelling of Vaccine Effect

The effect of the time-varying confounding variable age, in the SSCS method, has been modelled in three different ways as discussed in Chapter 2 and Chapter 5: (1) parametric SCCS where the age effect is included in the model as a step function, (2) semi-parametric SCCS where the age effect is left unspecified and (3) spline-based SCCS where the age-specific relative incidence is represented by a linear combination of cubic M-splines, whereas the effect of exposure to vaccines and other drugs is always represented by piecewise constant functions. The focus of this present chapter, therefore, is the representation of the relative incidence function associated with exposure within SCCS vaccine studies using flexible functions. This work is an extension to the standard SCCS method, recently proposed by Ghebremichael-Weldeselassie *et al.* (2014b). There has been much work on flexible ways of modelling the exposure effect for standard study designs (case control and cohort study designs). These involve representing the exposure history as a convolution of past exposures that combines information about duration, intensity and timing of exposure in one summary measure, as proposed by Breslow *et al.* (1983) and Thomas (1988). Letting $z(u)$ to be dose or intensity of exposure at time u and $w(u,t)$

a function that assigns weights to past exposures, the weighted cumulative exposure at time t is defined as

$$WCE(t) = \int_0^t z(u)w(u,t)du.$$

Within this context, interest has focused on modelling the weight function $w(u,t)$, whether by a priori chosen parametric models (Vacek, 1997; Langholz *et al.*, 1999; Abrahamowicz *et al.*, 2006) or spline models of varying complexity (Hauptmann *et al.*, 2000, 2001; Berhane *et al.*, 2008; Sylvestre and Abrahamowicz, 2009), with applications to environmental and drug exposures.

Four alternative parametric forms of the weight function were selected a priori by Vacek (1997) in a case control study design and the fits of the resulting models were compared to select the best fit to the data. Langholz *et al.* (1999), in the same study design, proposed to fit the weight function as parametric bilinear and exponential decay functions of time since exposure and applied it to a data on Colorado Plateau uranium miners to analyze latency effects of exposure to radon on lung cancer. Abrahamowicz *et al.* (2006) on the other hand proposed to use a priori selected parametric forms of the weight function or latency function to study the risk of fall related injuries among elderly new users of three benzodiazepines (nitrazepam, temazepam, and flurazepam) in a cohort study design using the Cox proportional hazards model.

In the case of vaccines, a point exposure occurs at the age of vaccination c, so $z(u)$ is a Dirac delta function. Setting $w(u,t) = w(t-u)$ we obtain the WCE function

$$WCE(t) = w(t-c) \text{ for } t > c, \text{ 0 otherwise.}$$

While our focus is on vaccines, the approach developed here has broader applicability, as will be shown in one of our examples in Section 6.3. In the standard SCCS

methodology, $WCE(t)$ (exposure-related relative incidence function) is represented by a step function, with pre-determined cut-points. This may not be biologically plausible and may incur losses in efficiency (Greenland, 1995a; Weinberg, 1995; Zhao and Kolonel, 1992). Furthermore, a poor choice of cut-points may be associated with cut-point bias and misclassification (Altman, 1991; Greenland, 1995b). We therefore propose a more flexible way of modelling the exposure effect in SCCS studies. We represent the exposure-related relative incidence function (which is a function of time since exposure or time since start of exposure in the context of drugs other than vaccines) as a linear combination of cubic M-spline basis functions described in Chapter 4.

The chapter is organized in four sections. In Section 6.1 we discuss representation of the exposure-related relative incidence function ($w(t-c)$) as a linear combination of cubic M-splines. Section 6.2 presents a simulation study conducted to evaluate the performance of the new method and compare it with the existing step function approach followed by application of the new approach, to febrile convulsions and MMR vaccine, and to fractures and thiazolidinedione use in Section 6.3. And finally in Section 6.4 we make some final remarks.

6.1 Spline-Based Exposure Risk Function

Regression splines provide smooth estimates with continuous first two derivatives and are flexible enough to represent a variety of clinically plausible shapes (Smith, 1979). Hauptmann *et al.* (2000) used constrained regression splines to represent the weight function in assessing the impact of exposure to smoking on lung cancer in a case control study. The weight function was represented as a linear combination of B-splines and the coefficients of the B-spline basis functions were constrained to be positive in order to obtain

non-negative weight function. Similarly Sylvestre and Abrahamowicz (2009) proposed the use of regression splines based on B-splines to model the weight function in cohort studies. The authors parameterized their model such that there will be no need for constrained optimisation. In Sylvestre and Abrahamowicz (2009) the weighted cumulative exposure is calculated at each time during follow up whereas in Hauptmann *et al.* (2000) it is evaluated only once at the end of follow up. Such extensions to model exposure effects were not introduced in the self-controlled case series method.

In the SCCS method we propose to approximate the exposure-related relative incidence function (the weight function) using regression splines based on M-splines. To begin with, we specify a nominal maximum risk period over which the exposure-related relative incidence function can be different from 1; outside this interval (which may be unbounded to the right), the function will take the value one. The argument of this function is time since start of exposure (in our case, vaccination). In the case of other drugs where the exposure is not a point exposure, there is no need to specify a nominal risk period, the duration of exposure is used as the risk period.

Recall that the likelihood function of SCCS derived in Chapter 2, for a single exposure, is given by:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\psi(t_{ij}) \exp\{x_i(t_{ij})\beta\}}{\int_{a_i}^{b_i} \psi(t) \exp\{x_i(t)\beta\}\, dt}. \tag{6.1}$$

where a_i and b_i are the start and end of observation period for individual i, t_{ij} is age of individual i when event j occurs, $\psi(t)$ is age-specific relative incidence function and $x_i(t)$ is exposure history of individual i (see Chapter 2 for more details). From the likelihood function (6.1), the exposure-related relative incidence function is required to be a positive function. Therefore, we use a linear combination of cubic M-spline basis functions, which are positive functions, to represent it. An M-spline of order q, as described in Chapter 4,

is a positive function constructed by combining pieces of polynomial functions of degree $q-1$ connected at knots. To keep positivity of the M-splines when combined linearly, we constrain their coefficients to be non-negative. Therefore, the function representing the exposure effect in equation 6.1, $\exp\{x_i(t_{ij})\beta\}$, will be replaced by a function of time since exposure represented as a linear combination of M-splines of order four:

$$\omega(t-c) = \begin{cases} \sum_{l=1}^{m} g(\beta_l) M_l(t-c), & c < t \leq d \\ 1, & \text{otherwise,} \end{cases}$$

where $g(\beta_l)$ are parameters to be estimated to determine the shape of the function, c is age at start of exposure, d is age at end of the nominal risk period and m is the number of M-spline functions. We shall choose $g(\beta_l) = \beta_l^2$ to ensure positivity of the function. The value m depends on the number of interior knots and the order of M-splines chosen: $m =$ number of interior knots + order. Usually a number of interior knots between 8 and 12 is sufficient (Joly *et al.*, 1998). We choose equidistant knots between 0 and maximum of $d_i - c_i$ (or the length of the nominal risk period for point exposures like vaccine), inclusive, and add an extra $q-1$ equidistant knots below the minimum and above the maximum knots to construct the M-spline basis functions. d_i is age at the end of exposure for individual i. When $d_i = \infty$ we set it equal to the value of b_i.

Replacing the exposure effect in Equation (6.1) by a linear combination of cubic M-splines gives the log-likelihood function

$$l = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log \left(\frac{\psi(t_{ij}) (\sum_{l=1}^{m} \beta_l^2 M_l(t_{ij} - c_i))^{I(c_i < t_{ij} \leq d_i)}}{\int_{a_i}^{b_i} \psi(t) (\sum_{l=1}^{m} \beta_l^2 M_l(t - c_i))^{I(c_i < t \leq d_i)} dt} \right), \tag{6.2}$$

where $I(c_i < t_{ij} \leq d_i)$ is an indicator variable that takes the value one if the event time is within the nominal risk period and zero otherwise.

The age-specific relative incidence is represented by a step function, as in the standard SCCS method. Thus, we subdivide the observation period of each case into intervals

$(l_{ih}, u_{ih}]$, h indexing the age group, with age-specific relative incidence $\exp(\alpha_h)$. Without loss of generality, we can choose these intervals to be sufficiently narrow (by splitting them) that they are properly contained in $(c_i, d_i]$ or its complement in $(a_i, b_i]$. The log-likelihood is then:

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{\exp(\alpha_{h(i,j)})(\sum_{l=1}^{m}\beta_l^2 M_l(t_{ij}-c_i))^{I(c_i<t_{ij}\leq d_i)}}{\sum_h \exp(\alpha_h)\int_{l_{ih}}^{u_{ih}}(\sum_{l=1}^{m}\beta_l^2 M_l(t-c_i))^{I(c_i\leq l_{ih}<d_i)}dt}\right). \tag{6.3}$$

where $h(i,j)$ is the age interval containing t_{ij}.

The integral in the denominator of the log-likelihood function (6.3) can be replaced by a linear combination of integrated splines (I-splines), since the integral of an M-spline function of order q can be expressed as an I-spline of order $q+1$. Hence, denoting the length of interval h for the i^{th} individual by $e_{ih} = u_{ih} - l_{ih}$, our log-likelihood function will be:

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(\frac{\exp(\alpha_{h(i,j)})(\sum_{l=1}^{m}\beta_l^2 M_l(t_{ij}-c_i))^{I(c_i\leq t_{ij}\leq d_i)}}{\sum \exp(\alpha_h)(e_{ih})^{(1-I(c_i\leq l_{ih}<d_i))}(\sum_{l=1}^{m}\beta_l^2 I_l(u_{ih}-c_i)-\sum_{l=1}^{m}\beta_l^2 I_l(l_{ih}-c_i))^{I(c_i\leq l_{ih}<d_i)}}\right). \tag{6.4}$$

To estimate the parameters of interest from the log-likelihood in (6.4), we introduce a roughness penalty term that controls the smoothness of the exposure-related relative incidence function. As in (O'Sullivan, 1988a) the penalty is based on the second derivative of the linear combination of cubic M-splines. Thus, the penalised log-likelihood function is:

$$\begin{aligned} pl &= l - \lambda \int \left(\sum_{l=1}^{m}\beta_l^2 M_l''(u)\right)^2 du \\ &= l - \lambda((\boldsymbol{\beta}_l^2)^T \mathbf{A}\boldsymbol{\beta}_l^2) \end{aligned} \tag{6.5}$$

where l is the log-likelihood function given in Equation 6.4, $\mathbf{A}$ is an $m \times m$ matrix with (r,l) element $\int M_r''(u)M_l''(u)du$ and $\lambda \geq 0$ is a smoothing parameter that controls the balance between smoothness of the function and fit to the data. One can also use a difference penalty as in (Eilers and Marx, 1996). The smoothing parameter λ is chosen by

maximising an approximate cross-validation score while keeping the age effect to be constant (i.e. setting the $\alpha_h = 0$). The approximate cross-validation score to be maximised is similar to the score in Equation 5.8 of Chapter 5 and is given as

$$\bar{V}(\lambda) = l(\hat{\boldsymbol{\beta}}) - \text{tr}([\hat{H} - 2\lambda\mathbf{S}]^{-1}\hat{H}), \tag{6.6}$$

since β_l^2 is used to keep positivity of the exposure relative incidence function

$$\mathbf{S} = 4\left(\mathbf{A}\text{o}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\right) + 2(\text{diag}(\mathbf{A}\boldsymbol{\beta}^2)).$$

Once the smoothing parameter is chosen, we maximise the penalised log-likelihood function (6.5) for fixed λ to estimate the parameters related to age and exposure effects.

6.1.1 Approximate Confidence Bands

In the context of cross-validated smoothing spline models, Wahba (1983) proposed Bayesian technique to generate confidence bands in which an improper prior for the function to be estimated was constructed using an integrated Wiener process. Silverman (1985) modified the idea of Wahba (1983) and came up with the same results using simpler and more intuitive priors. Silverman (1985) uses the roughness penalty term in a penalised log-likelihood function to be a prior log-likelihood. In simulation studies, Bayesian confidence intervals based on this approach proved to have good coverage properties, provided coverage is measured across the function, rather than point wise (Wood, 2006b). O'Sullivan (1988a) proposed the use of an approximate Bayesian technique for generating confidence bands for penalised likelihood estimators in the context of survival analysis, used in several applications including O'Sullivan (1988b) and Joly *et al.* (2002). Following O'Sullivan (1988a) we use the Bayesian-like technique to generate confidence bands for the exposure-related relative incidence estimators.

The penalised log-likelihood function (6.5) has two parts; the log-likelihood function (l) and the penalty component ($\lambda((\boldsymbol{\beta}_l^2)^T \mathbf{A} \boldsymbol{\beta}_l^2)$). Then considering the penalty term to be a prior log-likelihood for $\boldsymbol{\beta}$ leads in principle to the penalised log-likelihood function (6.5) to be a posterior log-likelihood for $\boldsymbol{\beta}$. Expanding the posterior log-likelihood in a second order Taylor series about the posterior mode of $\hat{\boldsymbol{\beta}}$ gives an approximate covariance of ($\hat{\boldsymbol{\beta}}$) to be $\hat{V}_{pl}$, where $\hat{V}_{pl}$ is the negative of the inverted hessian of the penalised log-likelihood function evaluated at the penalised maximum log-likelihood estimates.

Our approximation of the exposure-related relative incidence function used $g(\beta_l) = \beta_l^2$ to keep positivity of the function, we therefore need to know the covariance of β_{rl}^2. The required covariance matrix can be obtained using the delta method as

$$\hat{V}_{tr} = 4\text{diag}(\hat{\boldsymbol{\beta}})[\hat{V}_{pl}](\text{diag}(\hat{\boldsymbol{\beta}}))^T.$$

Hence an approximate 95% confidence interval for the exposure-related relative incidence at a point τ is

$$\hat{\omega}(\tau) \pm 1.96\sqrt{M(\tau)^T \hat{V}_{tr} M(\tau)}$$

where τ is time since first exposure and $M(\tau)^T = (M_1(\tau), \ldots, M_m(\tau))$.

Alternatively, to ensure that the confidence bands lie above zero, they can be obtained on the log scale as

$$\hat{\omega}(\tau) \exp\{\pm 1.96\sqrt{M(\tau)^T \hat{V}_{tr} M(\tau)}/\hat{\omega}(\tau)\}.$$

The confidence bands obtained in this way, however, are not really confidence bands for $\omega(\tau)$ rather they are confidence bands for $\bar{\omega}(\tau) = \mathbb{E}(\hat{\omega}(\tau))$, which can be taken as smoothed version of $\omega(\tau)$ (Wasserman, 2006). Therefore, we have to be cautious in reporting the results as the confidence interval will not be centered around the true function $\omega(\tau)$ due to the smoothing bias.

6.1.2 Fitting the Model

In order to fit the SCCS model with smooth exposure effect the data are required to be pre-processed based on age groups and the nominal risk period chosen, in a similar way to the parametric version of SCCS. The observation period of each event is subdivided into intervals and the data are reformatted such that each row contains information about a specific interval. The information contained in each interval after reformatting are: number of events (0 or 1), age at event (same for all intervals), upper limit, lower limit, length of the interval, age at start of exposure (same for all intervals of an event), age group and exposure status (a binary variable that indicates whether the specific interval lies within the control or the nominal risk period). For example suppose individual 1 who has been observed from age $0 - 730$ days experienced an event of interest at age 161 days and was vaccinated at 605 days of age. Then choosing the age cut points to be $426, 487, 548, 609$ and 670 days and a nominal risk period of 49 days the data for individual 1 are expanded as in Table 6.1

Table 6.1: *Data from a single event reformatted such that the observation period is divided based on age groups and a nominal risk period*

Indiv	Events	Eventday (t)	Upper limit (u)	lower limit (l)	Length (e)	Age at Exposure (c)	Age group	Exposure status
1	1	161	0	426	426	605	1	0
1	0	161	426	487	61	605	2	0
1	0	161	487	548	61	605	3	0
1	0	161	548	605	57	605	4	0
1	0	161	605	607	2	605	4	1
1	0	161	607	609	2	605	4	0
1	0	161	609	670	61	605	5	0
1	0	161	670	730	60	605	6	0

Once the data are reformatted the required ingredients of the log-likelihood function can be computed after selecting the number of knots that define the exposure-related relative incidence function. The function is defined between zero and the length of the nominal risk period or maximum of age at the end of exposure minus age at the start of exposure for non-point exposures. However, the time since exposure for events that occur before exposure is negative. Therefore, to compute the M-splines we replace the negative time since exposure values by zero and for events that occur beyond the nominal risk period we change their time since exposure value to the length of the nominal risk period. These changes will have no effect because the linear combination of the M-spline functions is forced to be one for the events occurring in the control periods by the indicator variable introduced in the log-likelihood function (6.4). And the I-splines in the denominator of the log-likelihood can be obtained in same way. After computing the M and I-splines at the required values, we choose the smoothing parameter using the approximate cross validation method by assuming no age effect. We then maximise the penalised log-likelihood function (6.5) fixing the smoothing parameter at the optimum value.

6.2 Simulation Study

To evaluate the performance of the new approach and compare it with the standard SCCS model, we conducted a simulation study. In this section we describe the design of the simulation study and results.

6.2.1 Design of the Simulation Study

We fixed the number of cases in all the data sets to be generated at 1,000. The length of the observation period for all cases was chosen to be 730 days, where age at the start of observation $a_i = 0$ days and age at the end of observation $b_i = 730$ days for all cases. Ages at vaccination or start of exposure (c_i) were uniformly distributed, see Section 7.4 of Chapter 7 for performance of the new method when c_i are not uniformly distributed.

Four different scenarios of true exposure-related relative incidence function were considered and generated from beta densities (Figure 6.1).

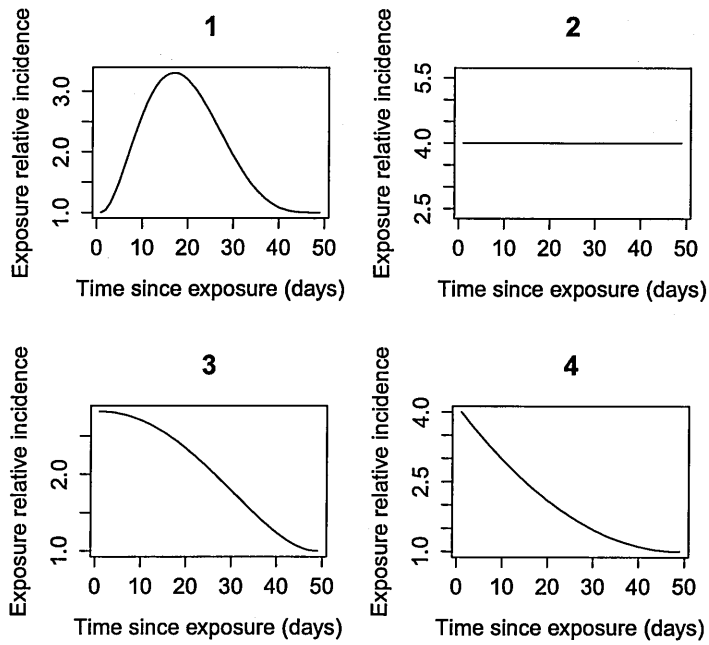

Figure 6.1: *True exposure-related relative incidence curves used in simulating the samples*

These functions show how the true exposure relative incidence values change with time since vaccination or start of exposure. The risk periods considered in all scenarios were of length 49 days. Hence 49 relative incidence values were generated based on the shapes chosen and obviously the true exposure relative incidence values outside the risk period

(in the control periods) were taken to be one. The effect of age was represented using a step function in which we used six age groups; $(0-426]$, $(426-487]$, $(487-548]$, $(548-609]$, $(609-670]$ and $(670-730]$ with true relative incidence rates $1, 2, 5, 8, 10$ and 15 respectively. Performance of the method developed in this chapter when the true age-specific relative incidence function is different from a step function (a sine function) can be seen from Section 7.4 of Chapter 7.

6.2.2 Data Generation

To generate number of events per individual and ages at event, first the observation period of each case is divided into intervals based on age and exposure groups. The intervals within the exposure risk period have length of one day because the true exposure relative incidence values change with age. Then incidence rates at each interval are computed as the age-specific incidence rate times the exposure-related relative incidence. Within the control periods, this is simply the relative incidence of the age group. From these, an average incidence rate for each individual is calculated to generate the marginal number of events per individual from a truncated Poisson distribution. Given the number of events for an individual generated from the truncated Poisson and incidence rates in each interval within the observation period, a multinomial distribution was used to identify in which interval the events occurred. Then a uniform distribution was used to generate event ages within the interval found to have an event. For each scenario 100 samples of 1,000 cases were generated in this way.

6.2.3 Analysis

Each of the simulated data sets were analysed using both the standard SCCS and the new approach with risk periods totalling 49 days following exposure (as simulated)

or with an extended nominal risk period of 98 days. In the standard SCCS, the risk period of 49 days following an exposure was divided in to seven groups of length seven days (with seven parameters). We also used an extended nominal risk period of 98 days, and fitted a standard SCCS model with 14 seven-day groups (and 14 parameters). In addition, we fitted the standard SCCS model with 49-day risk intervals (and hence one or two parameters, according to the nominal risk period). In all the spline-based analyses we used nine interior knots and the approximate cross-validation score was employed to choose the smoothing parameter. The standard SCCS method was fitted to evaluate the performance of the new method relative to it.

To compare the performance of the spline-based and standard SCCS methods we calculated the distances between the estimated and true exposure relative incidence functions. We used the Integrated Squared Error (ISE) to measure the distance, that is

$$\int_c^d (\omega(t-c) - \hat{\omega}(t-c))^2 dt,$$

where $\omega(t-c)$ is the true exposure-related relative incidence function, $\hat{\omega}(t-c)$ the estimated relative incidence function, c age at the start of exposure and d age at the end of exposure or nominal risk period. We computed the mean (MISE) and the standard deviation (SD) of the integrated squared error values obtained from the 100 samples. In addition to the MISE we used plots to make comparisons.

6.2.4 Results

In this section we present results of the simulation study. Figures 6.2 and 6.3 show the estimated exposure-related relative incidence curves obtained by fitting the spline-based and standard SCCS methods to the 100 randomly selected samples. The mean and standard deviation of the integrated squared errors are presented in Table 6.2.

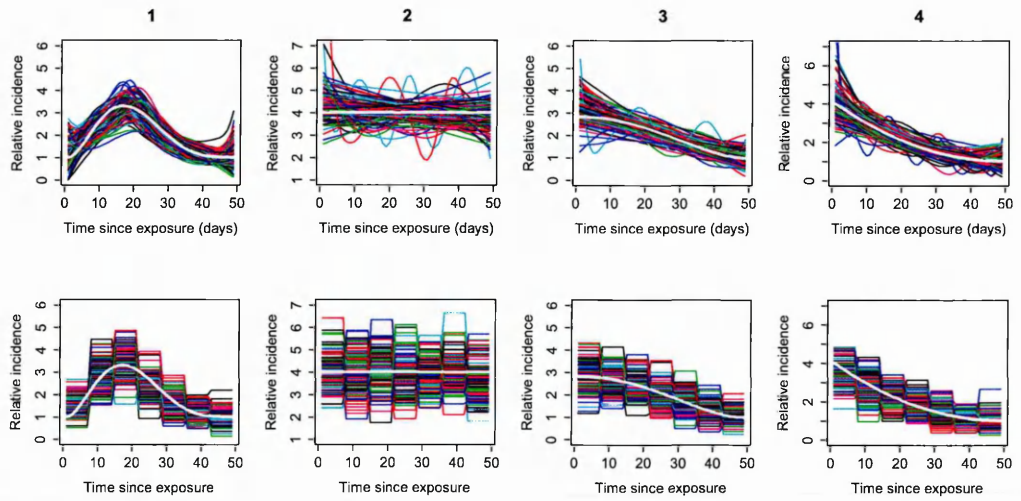

Figure 6.2: *Estimated relative incidence curves obtained by fitting the spline-based and standard SCCS to 100 randomly selected samples with the true relative incidence functions in thick white. Top row: estimates from the spline-based method; bottom row: results from the standard SCCS. Nominal risk period of 49 days was used.*

The results presented in Figure 6.2 are obtained when the risk period is kept at 49 days post exposure (which is equal to the risk period used to simulate the data). The top row presents results from the spline-based method and in the bottom row are results from the standard method. Results obtained by analysing the 100 randomly selected samples using both methods with a nominal risk period of 98 days are presented in Figure 6.3. The curves estimated from the standard method are step functions. The results from the spline-based analysis show that the shapes of the true relative incidence curves (thick white lines) were captured well by most of the estimated curves and all are included within the range of estimated curves in all scenarios. The variability of the estimated curves by the spline method is less as compared to the curves estimated by the standard SCCS method. Especially in scenario 2, where the true function is a constant, the variability from the standard method is very high. This is because we are estimating a constant

using seven or 14 parameters, which indicates loss of efficiency from the standard SCCS when we use a large number of exposure groups.

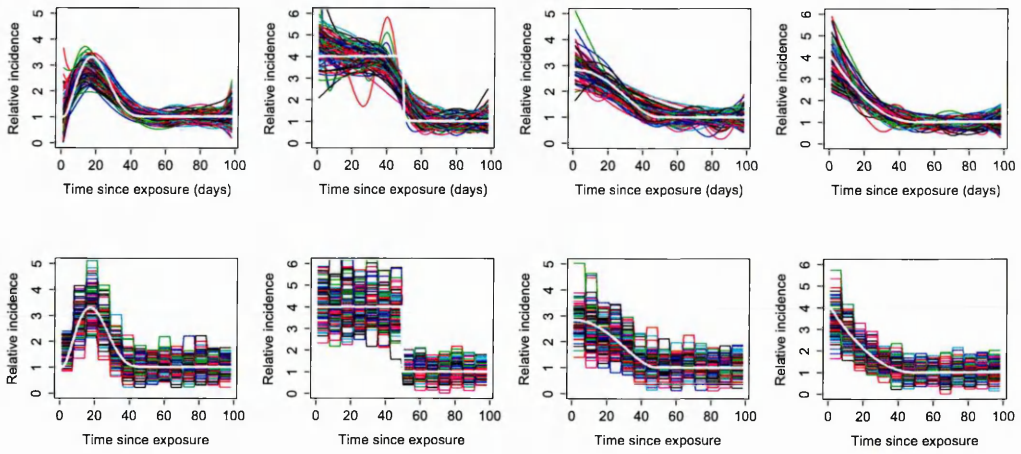

Figure 6.3: *Estimated relative incidence curves obtained from fitting spline-based and standard SCCS to 100 randomly selected samples with nominal risk period of 49 days. The thick white curve represents the true relative incidence function. Top row: estimates from spline method; bottom row: results from standard SCCS.*

Similar to the results from the spline method, the true relative incidence curves lie within the estimated curves obtained from fitting the standard SCCS method.

Table 6.2 presents MISE and and standard deviation of the integrated squared errors for all the four exposure-related relative incidence function scenarios presented in Figure 6.1.

Table 6.2: *Mean integrated squared error (MISE) and standard deviation (SD) obtained from spline-based and standard SCCS models. Each simulated data set was fitted twice by the two methods with nominal risk periods of 49 and 98 days*

	Spline-based SCCS		Standard SCCS with groups of length 7 days		Standard SCCS with groups of length 49 days	
Scenario	MISE	SD	MISE	SD	MISE	SD
			Potential risk length of 49 days			
1	7.982	5.685	14.993	8.202	37.934	3.494
2	9.575	10.190	31.368	16.434	5.498	7.559
3	5.453	5.625	12.338	6.207	22.388	2.924
4	6.478	8.376	14.650	7.300	43.490	4.593
			Potential risk length of 98 days			
1	14.875	7.096	20.072	7.926	38.121	3.414
2	34.112	13.747	38.750	18.549	8.012	10.127
3	6.439	5.283	20.000	18.791	22.654	2.659
4	8.151	6.823	19.037	8.201	44.232	3.059

Table 6.2 shows that the mean integrated squared errors (MISE) are all lower for the spline method than the standard method, except for scenario 2, in which the true exposure-related relative incidence was constant. For this scenario, the correctly specified step function model (with one or two parameters) outperforms the spline model, though interestingly, the over-specified step function model (with seven or 14 parameters) does not. Comparable but slightly degraded results were obtained for scenarios 1, 3 and 4 with the 98-day nominal risk period as with the correct 49-day risk period. For scenario 2, the spline method produced worse results with the 98 day risk period compared to 49 day nominal risk period.

Figures 6.4 and 6.5 present the systematic error or bias at a point τ (time since exposure) and standard deviation of estimated exposure relative incidence values at a

point τ. The bias was calculated as

$$average(\hat{\omega}(\tau)) - \omega(\tau),$$

where $\omega(\tau)$ is the true relative incidence at point τ, $\hat{\omega}(\tau)$ estimated relative incidence at point τ and the average is the mean of the estimated relative incidences at point τ over the 100 runs of simulations. The standard deviation is taken over the 100 $\hat{\omega}(\tau)$ values.

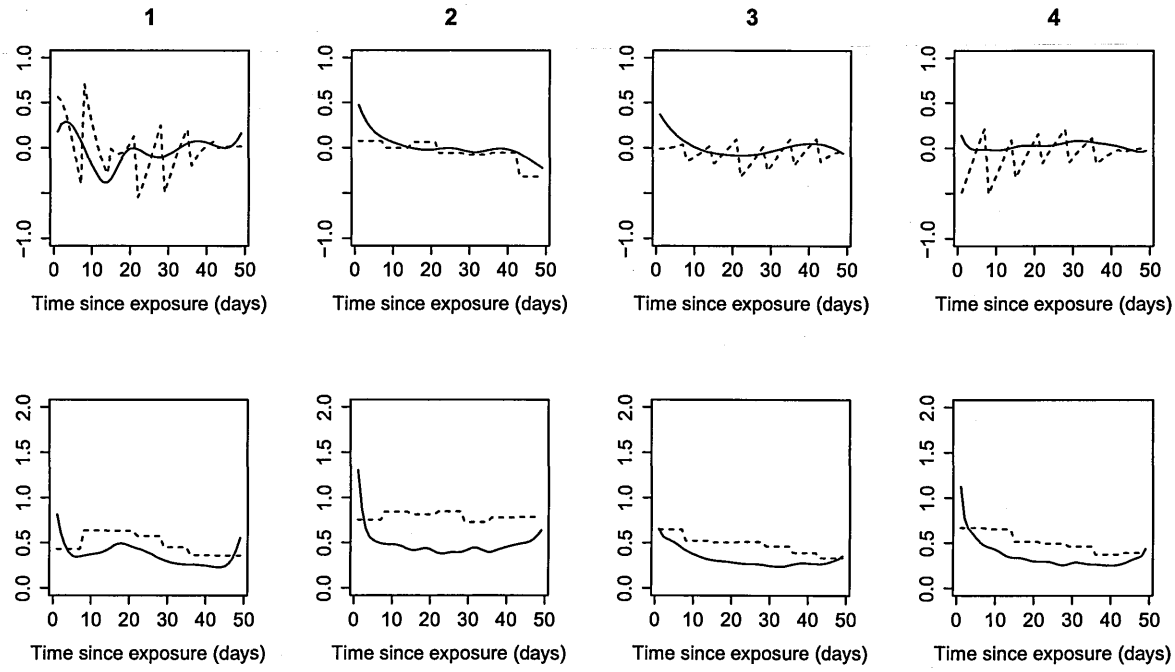

Figure 6.4: *Bias (top row) and standard deviation (bottom row) of estimates obtained by fitting the spline-based SCCS (solid lines) and the standard SCCS (dotted lines) with nominal risk period of 49 days to the simulated data sets. 7 exposure groups were used when fitting the standard SCCS.*

Figure 6.4 shows the bias (top row) and variability (standard deviation, bottom row) of estimates from the standard (with 7 parameters) and the spline-based SCCS methods with a 49 day post exposure nominal risk period. The bias of the standard method has a saw-tooth appearance in scenarios 1, 3 and 4 related to discontinuities at the cut-

points, whereas the spline method occasionally shows some bias at endpoints, notably for scenarios 2 and 3. The spline method produces lower standard deviations, except at the endpoints.

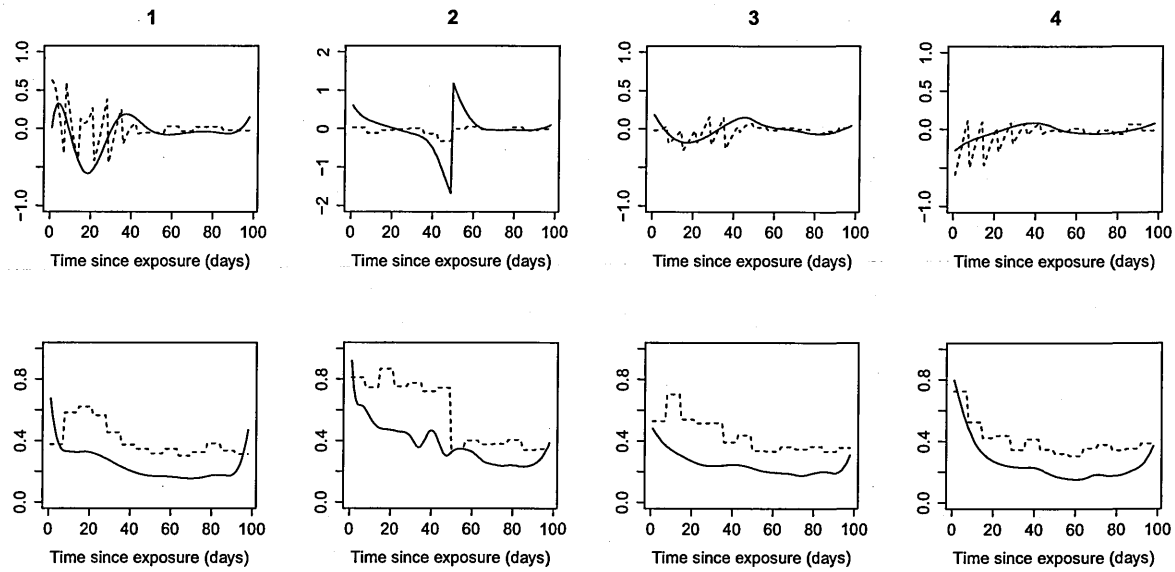

Figure 6.5: *Bias (top row) and standard deviation (bottom row) of estimates obtained by fitting spline-based SCCS (solid lines) and standard SCCS (dotted lines) to the simulated data sets. A nominal risk period of 98 days was used, divided into* 14 *exposure groups when fitting the standard SCCS.*

Similar results were obtained when the nominal risk period was extended to 98 days post exposure, as presented in Figure 6.5. However, for scenario 2 there is higher absolute bias for splines than the standard method around 49 days since exposure where the true relative incidence value of 4 drops to 1. The variability is still higher for the standard method than the spline method.

6.3 Application

In this section, we illustrate the use of the new method that represents vaccine effect using a smooth function (a linear combination of M-splines) and the standard SCCS method, for comparison purposes, to two data sets. The first application is on a data set of MMR vaccine effects introduced in Section 5.4 of Chapter 5. Although the method developed here focuses on representation of vaccine effects, it can be applied to non vaccine exposures and therefore we apply it to data on the effect of thiazolidinediones use in causing fracture.

6.3.1 Analysis of Febrile Convulsion Data

The aim of this analysis is to investigate a potential association between febrile convulsions and measles/mumps/rubella (MMR) vaccine using the new spline-based and standard SCCS methods. The data set, as described in Chapter 5, comprises of 2,389 children aged between 29 and 730 days in the period 1991–1994. They experienced 3,826 febrile convulsion events in total, indicating that there were children with more than one event of febrile convulsion. In this example, we used 50 days post MMR vaccine as a nominal risk period for all cases to represent the exposure effect with splines. Since all individuals have the same nominal risk period of 50 days, we defined 12 equidistant inner knots between 0 and 50 days. Age was included in the model as a step function. There were 21 age groups of length 30 days while the first and last groups were of length 32 and 40 days respectively.

A linear combination of cubic M-splines was used to represent the MMR-related relative incidence function. The value of the smoothing parameter selected by the approximate cross-validation score was 0.031. We present the relative incidence function

estimated by maximising the penalised log-likelihood function (6.5) along with its approximated confidence bands in Figure 6.6. The figure shows no risk of febrile convulsion in the first three days post MMR vaccination and a borderline non-significant relative incidence of 1.248 at the fourth day. However, there is a significantly increased risk between five and 11 days after exposure to the vaccine. The relative incidence at the 5th day is 1.922 and increases smoothly to 3.647 at the eighth day and then the risk decreases to 1.244 at 12 days since exposure. There is also an increased risk of febrile convulsion due to MMR vaccine between 19 and 21 days post vaccination. At all other times after vaccination there is no significantly increased risk of febrile convulsion.

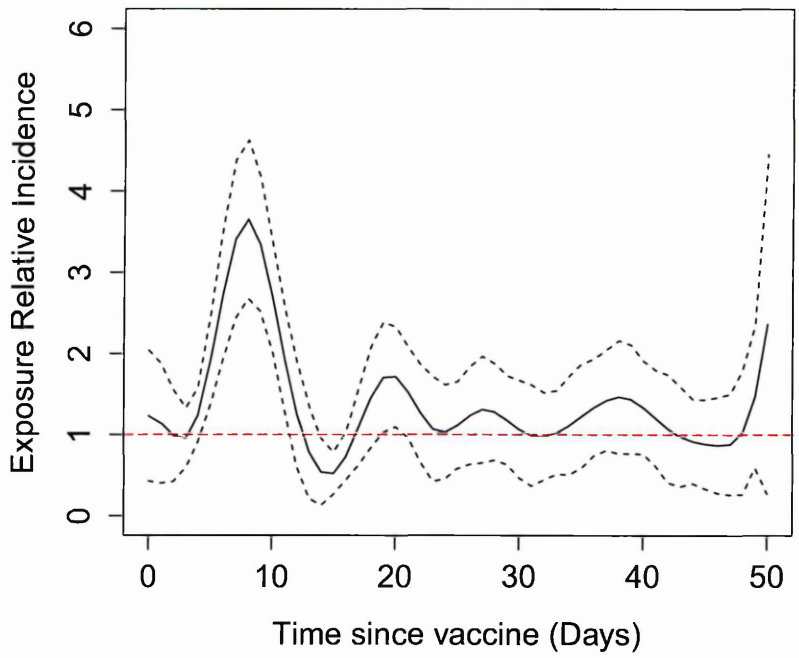

Figure 6.6: *Smooth estimate of the relative incidence function related to exposure to MMR vaccine (bold line) and 95% confidence bands(doted lines).*

We also fitted the standard SCCS method, where age and exposure effects are rep-

resented by piecewise constant functions, to the data set on febrile convulsion. We used the same 21 age groups described above to model the age effect and for the exposure effect we divided the 50 days post MMR vaccine nominal risk period in to 10 groups. The ten exposure groups had cut points at $6, 11, 18, 22, 26, 30, 36, 40$ and 45 days since vaccine. The results from this analysis are presented in Table 6.3 and Figure 6.7. Figure 6.7 presents the exposure to MMR vaccine specific relative incidence functions estimated from the standard SCCS model (step function) and the spline-based SCCS method (smooth function).

Table 6.3: *Relative incidence (RI) estimates of exposure to MMR vaccine and lower and upper 95% confidence intervals obtained from fitting parametric SCCS method with 10 exposure groups and 21 age groups*

Exposure Group	Relative Incidence	95% Confidence Interval	
(Days)	(RI)	Upper	Lower
0-6	1.226	0.922	1.629
6-11	3.489	2.881	4.225
11-18	0.827	0.600	1.139
18-22	1.493	1.087	2.050
22-26	1.089	0.752	1.577
26-30	1.167	0.816	1.671
30-36	1.159	0.862	1.559
36-40	1.368	0.980	1.909
40-45	0.977	0.687	1.390
45-50	1.103	0.790	1.539

Figure 6.7 shows that the results obtained from the two methods are similar since the exposure groups used in the standard method are correctly specified. However, different categorizations may give different results for the standard SCCS method, which is a disadvantage of the method if the correct exposure groups are not known.

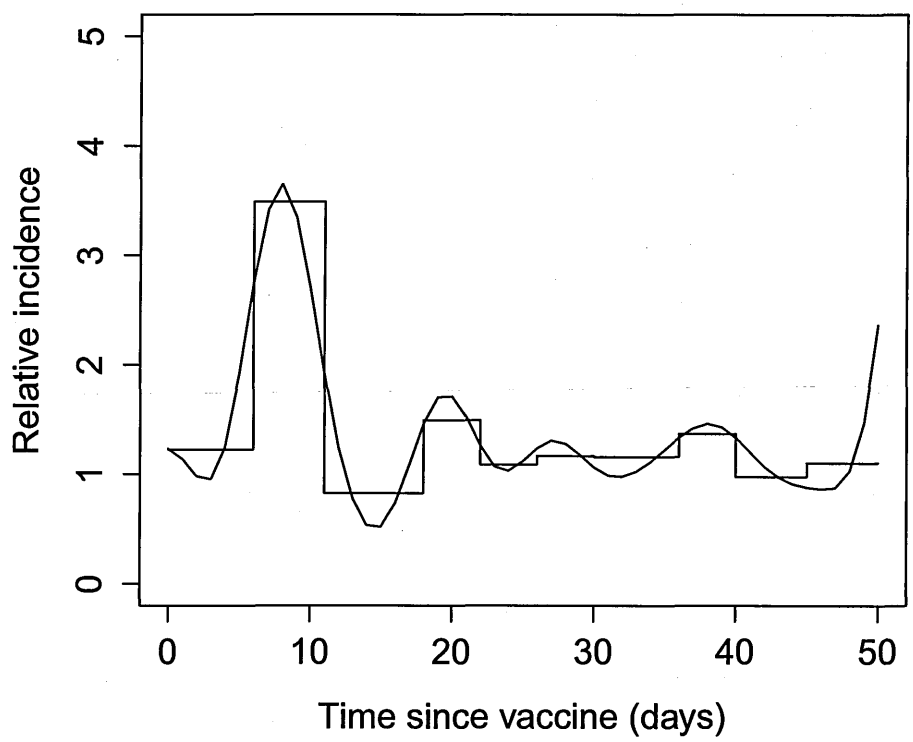

Figure 6.7: *Relative incidence functions related to MMR vaccine estimated from fitting the standard model with* 10 *exposure groups (step function) and spline-based SCCS (smooth function).*

6.3.2 Analysis of Fracture Data

The methods developed in the present chapter can be applied more widely. We illustrate this with data on fractures and thiazolidinediones, which were analysed by Douglas *et al.* (2009) using the standard case series method. Thiazolidinediones are a class of medicines used to treat type 2 diabetes.

Diabetes is a lifelong condition that causes a person's blood sugar level to become too high. In the UK, approximately 2.9 million people are affected by diabetes. There are also thought to be around 850,000 people with undiagnosed diabetes (NHS, 2013). There

are two main types of diabetes, referred to as type 1 and type 2. Type 2 diabetes occurs when the body does not produce enough insulin to function properly, or the body's cells do not react to insulin. This is known as insulin resistance. Type 2 diabetes is far more common than type 1 diabetes. In the UK, about 90% of all adults with diabetes have type 2 diabetes (NHS, 2013). Type 2 diabetes usually affects people over the age of 40, although increasingly younger people are also being affected.

Diabetes cannot be cured, but treatment aims to keep blood glucose levels as normal as possible to control symptoms and minimise health problems developing later. In some cases of type 2 diabetes, it may be possible to control symptoms by altering your lifestyle, such as eating a healthy diet and exercise. However, as type 2 diabetes is a progressive condition, it may eventually be needed to take medication to keep blood glucose at normal levels. To start with the medication usually takes the form of tablets, but later on it may include injected therapies, such as insulin.

There are different types of medicines recommended to treat diabetes 2, including metformin, sulphonylureas, thiazolidinediones. Thiazolidinediones also known as glitazones, make body cells more sensitive to insulin so that more glucose is taken from blood. The first type of thiazolidinediones, pioglitazone, is usually used in combination with metformin or sulphonylureas, or both. They may cause weight gain and ankle swelling (NHS, 2013). Another thiazolidinedione, rosiglitazone, was withdrawn from use in 2010 due to an increased risk of cardiovascular disorders, including heart attack and heart failure.

The aim of the study in Douglas *et al.* (2009) was to investigate whether there is an increased risk of fracture associated with the use of thiazolidinediones. The fractures considered were classified according to fracture site (ankle, arm, chest/rib, face, hand, hip, leg, pelvis, shoulder, skull, wrist, spine, multiple sites and unknown site).

The data used in the analysis were primary care computerized clinical records from the United Kingdom-based General Practice Research Database (GPRD). 1,819 patients aged about 40 years or older prescribed at least one thiazolidinedione and with at least one fracture event were included in the analysis. The data included patients with multiple fractures: 283 (16%), 64 (4%), and 25 (1%) had two, three, and four or more fractures, respectively. Multiple fractures were included in the analysis if the fractures happened at different sites or at the same site but at least 6 months apart. Out of the 1,819 patients 990 (54%) were women with mean age at first thiazolidinediones prescription of 65.4 years and mean age for men was 57.9 years.

In Douglas *et al.* (2009), the authors defined the control period to be from start of observation period until first prescription of a thiazolidinedione and the risk period was from age at start of thiazolidinedione use until age at end of observation period. The length of exposure following each individual prescription was calculated using information recorded in the GPRD on pack size and dosing frequency. Thiazolidinedione treatment was assumed to be continuous where any apparent treatment break was less than 60 days, to allow for partial noncompliance and situations where patients may have built up treatment stocks (Douglas *et al.*, 2009). Age at end of observation was then taken to be age at the earliest of any treatment break longer than 60 days or the end of recorded follow up in the database. The mean duration of control periods prior to thiazolidinedione use was 9.5 years, and the mean duration of exposure to a thiazolidinedione was 2.3 years.

Different analyses, using the standard SCCS method, were done by Douglas *et al.* (2009) including analysis for all fracture sites together with any type of thiazolidinediones (pioglitazone or rosiglitazone) exposure, for females only, males only, analysis by fracture site, analysis by taking patients who were exposed only to one type of thiazolidinediones

etc. Here we present only one of their analyses with all fractures together and exposure to any thiazolidinedione use and reanalyse the same using the new method of representing the exposure effect.

Unlike vaccines, thiazoledinediones are not point exposures, however we can use a similar approach as with vaccines by taking $z(u) = z$ for $u > c$, the age at first thiazeledinedione, so $WCE(t) = z\int_c^y w(t-u)du$ and in the SCCS context it will be $WCE(t) = zw(t-u)$. In the SCCS likelihood function the value z is cancelled out similar to the baseline incidence. This leads to the same likelihood function. We reanalyzed the data using the new version of SCCS where time since exposure is represented by a linear combination of M-splines. In our analysis, we used the same exposure risk periods as in Douglas *et al.* (2009). The maximum duration of exposure to thiazolidinedione was 2,364 days. Hence our exposure-related relative incidence function was represented by a linear combination of cubic M-splines defined between 0 and 2,364 days since first exposure. We chose 14 equidistant knots between 0 and 2,364 days inclusive, i.e we have 16 M-spline basis functions. The time-varying confounding covariate age was taken into account using a piecewise constant function with 42 age groups: the first age group is less than 14,610 days (40 years) of age, followed by five age groups of length two years, 28 groups of one year length, seven groups of length two years and the last age group with age greater than 33,603 days (92 years).

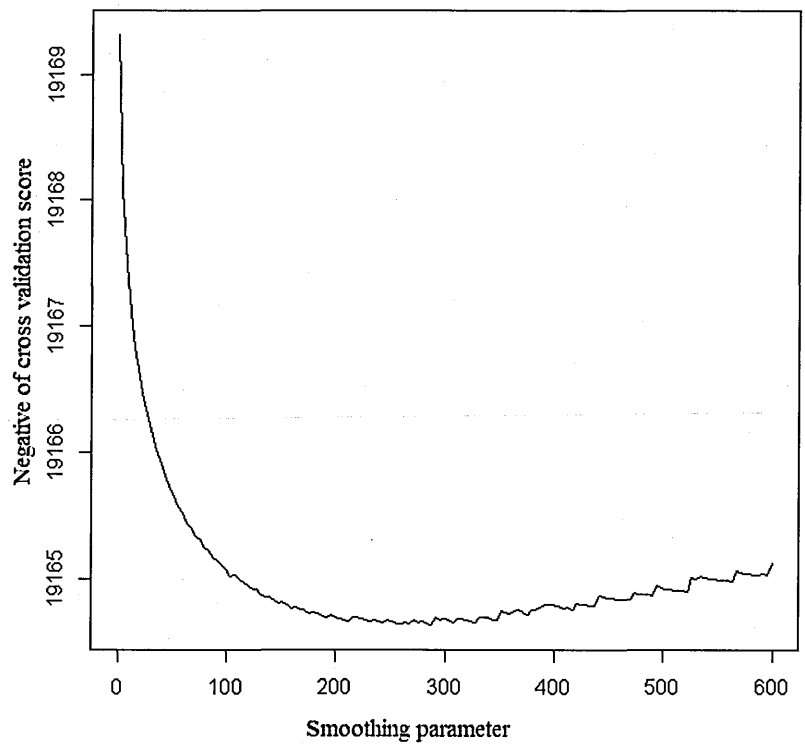

Figure 6.8: *Negative of the approximate cross validation score versus the smoothing parameter to choose the value of the smoothing parameter that maximises the approximate cross validation score.*

To estimate the parameters of interest, we first selected the optimum smoothing parameter, λ, that maximises the approximate cross-validation score in Equation (6.6). This optimum λ was 288 (Figure 6.8). Figure 6.8 plots the grid of smoothing parameter values, λ, versus the negative of the approximate cross validation score and shows that the optimum value for the smoothing parameter is 288. We then maximised the penalised log-likelihood function in Equation (6.5) for fixed $\lambda = 288$ to get the required parameters. The estimated exposure-related relative incidence function and its approximate confidence bands are presented in Figure 6.9.

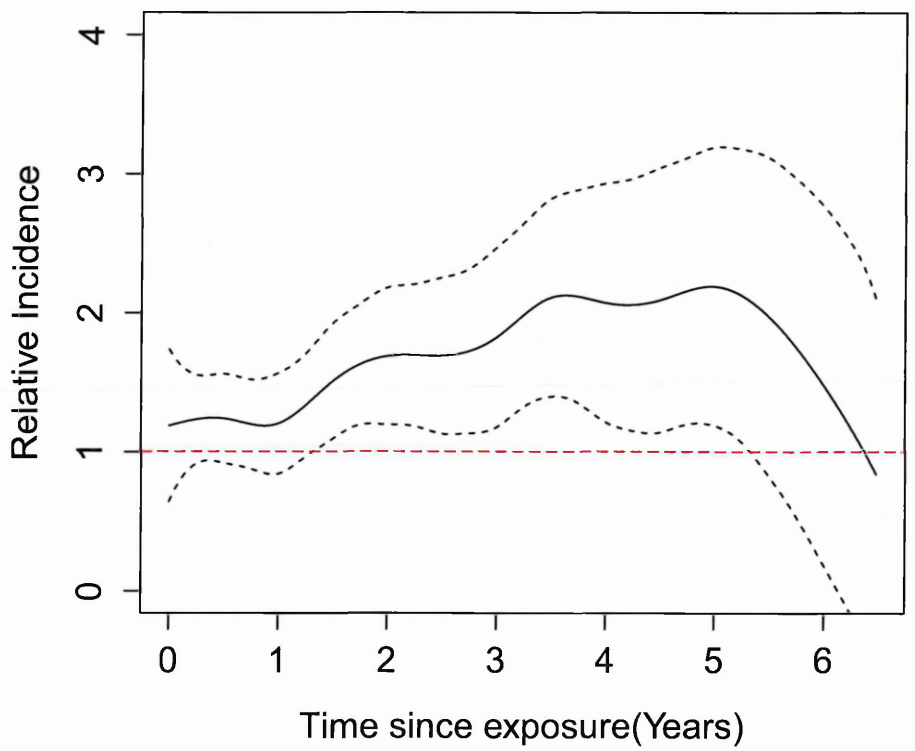

Figure 6.9: *Relative incidence function estimate related to thiazolidinedione use (bold line) and 95% confidence intervals (dotted lines).*

From Figure 6.9, it can be seen that the relative incidence of fracture due to thiazolidinedione use increases as time since exposure increases. There is no significant increased risk of fracture in the first two months of exposure and the relative incidence is borderline significant from two months to about one year and half, but there is a significantly increased risk of fracture due to exposure to thiazolidinedione thereafter, and the maximum relative incidence of 2.103 is reached after about 5 years of exposure. The relative incidence may start to decrease and the confidence bands widen after 5 years.

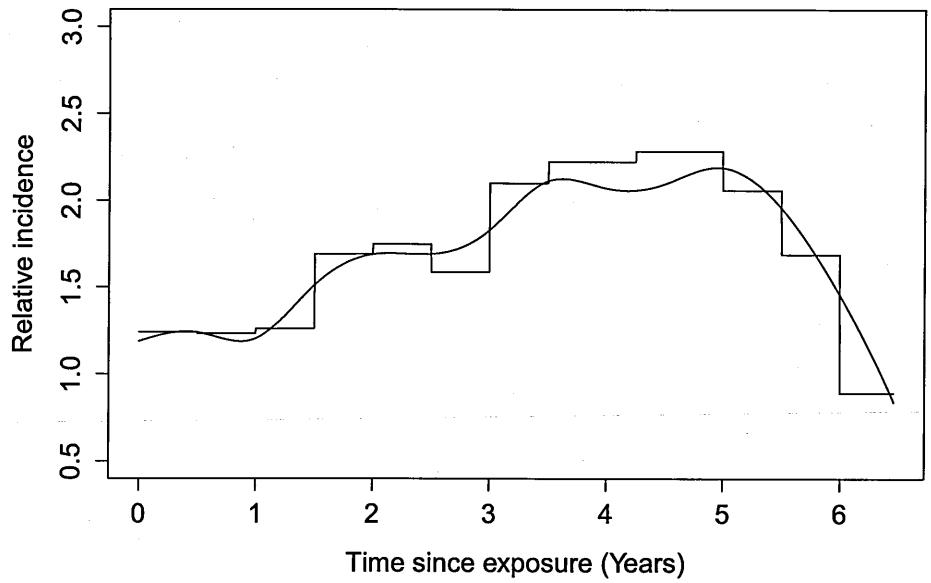

Figure 6.10: *Relative incidence functions related to thiazolidinedione use estimated by fitting the standard SSCS model with* 13 *exposure groups (step function) and the spline-based SCCS (smooth function).*

In their standard SCCS analysis, Douglas *et al.* (2009), defined five exposure groups of $(0-1), (1-2), (2-3), (3-4)$ and $(4-7)$ years since first exposure and obtained relative incidence estimates of $1.26, 1.49, 1.70, 2.31$, and 2.00 respectively. We repeated the analysis but with a different number and length of exposure groups. We divided the time since first exposure in to 12 groups of lengths six to nine months. Results from this analysis are presented in Figure 6.10 and Table 6.4. The results obtained from the standard SCCS method with 12 exposure groups are similar to those obtained by the spline method.

Table 6.4: *Relative incidence (RI) estimates of exposure to thiazolidinedione and lower and upper 95% confidence intervals obtained from fitting parametric SCCS method with 12 exposure groups and 42 age groups*

Exposure Group	Relative Incidence	95% Confidence Interval	
(Years)	(RI)	Upper	Lower
0.0 - 0.5	1.242	1.024	1.506
0.5 - 1.0	1.233	0.996	1.527
1.0 - 1.5	1.262	0.997	1.596
1.5 - 2.0	1.691	1.335	2.143
2.0 - 2.5	1.748	1.346	2.269
2.5 - 3.0	1.587	1.174	2.146
3.0 - 3.5	2.099	1.546	2.850
3.5 - 4.25	2.223	1.646	3.003
4.25 - 5.0	2.284	1.591	3.280
5.0 - 5.5	2.059	1.183	3.583
5.5 - 6.0	1.690	0.770	3.709
6.0 - 7.0	0.894	0.122	6.535

6.4 Discussion

In this chapter, we proposed using penalised regression splines to model the effect of point exposures due to vaccination, and drug-related exposures more widely, in the self-controlled case series method. We model the exposure-related relative incidence function as a linear combination of cubic M-splines. This approach avoids the limitations of the standard and semi-parametric SCCS methods that use step functions with pre-specified cut-points to assess the exposure effect.

Our spline-based SCCS method can be considered as a special case of weighted cumulative exposure models used in environmental epidemiology, which have also made good use of spline models (Hauptmann *et al.*, 2000; Sylvestre and Abrahamowicz, 2009). These

approaches have used information criteria to choose the number of knots in defining the B-spline basis functions. In our case, we intentionally selected a large number of knots and introduced a penalty term to the log-likelihood function to avoid over-fitting, the smoothing parameter being chosen by an approximate cross validation score (O'Sullivan, 1988a; Joly *et al.*, 1998, 2002). An approximate Bayesian like method was used to produce confidence bands for the exposure-related relative incidence function. However, this method does not take the variability due to choosing the smoothing parameter into account. Bootstrapping is another option to generate the confidence bands but is computationally expensive.

Simulation studies showed that the new approach generally has a better performance than the use of step functions in the context of the SCCS method. The new method was applied to two data sets to investigate the association between febrile convulsions and MMR, and between fracture and thiazolidinedione use. The estimates obtained from the new method are consistent with the results from the standard SCCS method when the exposure groups are correctly specified. Increasing the number of a priori defined exposure groups in a standard SCCS model may help in capturing the true exposure-related relative incidence curve better, but at the cost of reduced efficiency. The new method is likely to be particularly useful in the absence of a clear, a priori hypothesis regarding the risk period. It can also be used to obtain an overall risk profile, or, if required, to specify risk periods upon which to base standard SCCS analyses in other data sets.

While our focus has been on developing methods for studying the safety of vaccines, they have wider applicability, as we have shown in our example on fractures and thiazolidinediones. In the example on fractures and thiazolidinediones, we showed an application of the new method when the exposure period was from age at first prescription of the drug

until the end of the observation period when there is no control period after the end of exposure. In addition, the method can be applied when there are interrupted exposures, that is individuals will have different length of exposure periods and the observation period goes beyond the end of exposure. This is done by assuming that the periods before the start of exposure and immediately after the end of exposure up to the end of observation period are control periods. However, in many pharmacoepidemiological studies it may be necessary to study the effect of exposure in the wash-out period, a period immediately after the end of exposure to drug. In this respect the new SCCS approach needs further extension.

Chapter 7

Non-Parametric Self-Controlled Case Series Method

In Chapter 5, to avoid the limitations of the standard and semi-parametric versions of the SCCS method in modelling the age effects we represented the age-specific relative incidence by a linear combination of M-spline functions. While the age effect was represented by a smooth function (based on splines), the effect of exposure was modelled using a step function. In Chapter 6, instead of using a step function we proposed using a linear combination of M-splines to model the effect of exposure, as a function of time since exposure. However, the age effect was represented by a step function. Both these extensions to the standard SCCS method involve step functions. Therefore, in this chapter we propose modelling both age and exposure effects using splines to create a fully non-parametric extension to the SCCS method. After some initial remarks in Section 7.1, the likelihood function of the non-parametric SCCS method is derived in Section 7.2. In this section, we also describe and define derivatives and integrals of M and I splines, and the integral of a product of two spline functions. Section 7.3 presents the penalised log-likelihood function

of the non-parametric SCCS method and discusses the selection of smoothing parameters. In Section 7.4, we evaluate the performance of the new method using simulations. We apply the non-parametric SCCS method to data on febrile convulsion and MMR vaccine in Section 7.5 and finally follow this with a discussion in Section 7.6.

7.1 Modelling Age and Exposure Effects Using Splines

The use of regression splines in the context of the self-controlled case series method has shown an improved performance compared to the use of step functions as presented in Chapters 5 and 6. Among the motivations for using regression splines based on M-splines in these chapters were that the spline functions give flexible and plausible shapes of age and exposure-related relative incidence functions and avoid numerical integration of the integral in the denominator of the SCCS likelihood function. This numerical integration is avoided because the integral of an M-spline is an I-spline, therefore the integral of a linear combination of M-splines can be expressed as a linear combination of I-splines. Based on similar arguments, both age and exposure effects can be represented as linear combinations of M-spline basis functions. In this chapter, since age and exposure are to be represented by linear combinations of M-splines at the same time, the denominator of the SCCS likelihood function involves the integral of a product of two spline functions. This cannot be represented by a linear combination of I-splines only, so the integration cannot be avoided in the same way. Therefore, based on the definition of the integral of an M-spline developed by Ramsay (1988), we define first, second and third integrals of an I-spline. In the following section we derive the likelihood function of the SCCS method when both age and exposure effects are approximated by linear combinations of M-spline basis functions.

7.2 Likelihood Function

To derive the likelihood function of the non-parametric SCCS method, we begin with the general SCCS likelihood function derived in Chapter 2, Equation 2.9 and, for one exposure, given as

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\psi(t_{ij}) \exp\{x_i(t_{ij})\beta\}}{\int_{a_i}^{b_i} \psi(t) \exp\{x_i(t)\beta\}\, dt}$$

which in Chapter 6, we generalized to

$$L = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\psi(t_{ij})\omega(t_{ij} - c_i)}{\int_{a_i}^{b_i} \psi(t)\omega(t - c_i)dt} \tag{7.1}$$

where a_i and b_i are the start and end of the observation period for individual i, $\psi(t)$ is the age-related relative incidence function, $\omega(t-c)$ is the exposure-related relative incidence function which takes the value one if the event day is not between age at start of exposure (c_i) and age at end of exposure (d_i). In the standard SCCS method, $\psi(t)$ and $\omega(t-c)$ are represented by step functions; in the semi-parametric version of SCCS, $\psi(t)$ is left unspecified and $\omega(t-c)$ is fitted as a step function; in Chapter 5, $\psi(t)$ was approximated by splines and $\omega(t-c)$ by a step function, and in Chapter 6, $\psi(t)$ was represented as a step function and $\omega(t-c)$ as a linear combination of M-spline functions. In this chapter, we approximate both $\psi(t)$ and $\omega(t-c)$ as linear combinations of cubic M-spline basis functions.

As in Chapter 5, $\psi(t)$ is defined between $a = \min\{a_i; i = 1, \ldots, N\}$ and $b = \max\{b_i; i = 1, \ldots, N\}$, where N is the total number of cases in the study. Since $\psi(t)$ is a relative effect it has to be a positive function and to get such a function based on M-splines we constrain the coefficients to be non-negative and get the expression for $\psi(t)$ as in Equation 5.2

$$\psi(t) = \sum_{l=1}^{m_1} g(\alpha_l) M_{1l}(t) = \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t). \tag{7.2}$$

The $g(\alpha_l)$ are parameters used to determine the shape of $\psi(t)$ and are constrained to be non-negative by taking $g(\alpha_l) = \alpha_l^2$. $M_{1l}(t)$ is the l^{th} M-spline basis function related to age, m_1 is the number of parameters or the number of M-spline basis functions which is equal to the sum of the number of interior knots and the order of the basis functions.

Similarly, as in Chapter 6, an exposure-related relative incidence function with non-negative coefficients is defined between 0 and $\max\{(d_i - c_i); i = 1, \ldots, N\}$, where c_i and d_i are the start and end of age at exposure respectively for individual i. When the exposure is a point exposure, e.g a vaccine, a nominal risk period is defined which can be unbounded to the right. The nominal risk period is a period within the observation period where the exposure-related relative incidence can be different from 1 and outside it the exposure-related relative incidence function takes the value 1. Therefore, it is defined as:

$$\omega(t-c) = \begin{cases} \sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t-c), & c < t \leq d \\ 1, & \text{otherwise,} \end{cases} \tag{7.3}$$

where m_2 is the number of M-spline basis functions used to define the exposure-related relative incidence function, $\omega(t-c)$ and $M_{2l}(t-c)$ is the l^{th} basis function related to exposure. In this Chapter the knots which are used to define the M-splines related to the age effect and the exposure effect are chosen to be equidistant including the arbitrary knots added below and above the minimum and maximum values of the variable.

Now replacing $\psi(t)$ and $\omega(t-c)$ in Equation (7.1) by the spline functions in Equations (7.2) and (7.3) respectively gives the likelihood function for the non-parametric SCCS as

$$l = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{\left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t_{ij})\right) \left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t_{ij} - c_i)\right)^{I(c_i < t_{ij} \leq d_i)}}{\int_{a_i}^{b_i} \left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)\right) \left(\sum_{l=1}^{m} \beta_l^2 M_{2l}(t - c_i)\right)^{I(c_i < t \leq d_i)} dt}$$

and the log-likelihood function is

$$l = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log \left(\frac{\left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t_{ij})\right) \left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t_{ij} - c_i)\right)^{I(c_i < t_{ij} \leq d_i)}}{\int_{a_i}^{b_i} \left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)\right) \left(\sum_{l=1}^{m} \beta_l^2 M_{2l}(t - c_i)\right)^{I(c_i < t \leq d_i)} dt} \right). \tag{7.4}$$

To further simplify the denominator of the log-likelihood function (7.4) (so that it avoids numerical integration), we will use integration by parts. This will involve derivatives and integrals of linear combinations of M-spline functions and integrals of the integrals. Therefore, before we proceed with simplifying the log-likelihood function, we describe derivatives of M-splines and define integrals of I-splines in the following subsections.

7.2.1 Derivatives of M-splines

From Chapter 4, we have that M-splines of order q are defined as divided differences of truncated power functions, that is, for a given knot sequence $k_1 = k_2 = \cdots = k_q < k_{q+1} < \cdots < k_{q+s} < k_{q+s+1} = k_{q+s+2} = \cdots = k_{2q+s}$

$$M_l(t|q) = \frac{q}{k_{l+q} - k_l} B_l(t|q) = (-1)^q q [k_l, \ldots, k_{l+q}] T_t^q(k)$$

where $T_t^q(k)$ is a truncated power function of order q given by $T_t^q(k) = (t-k)_+^{q-1}$. Therefore, the first derivative of an M-spline function is

$$\frac{dM_l(t|q)}{dt} = (-1)^q q [k_l, \ldots, k_{l+q}] \frac{dT_t^q(k)}{dt}$$

and the derivative of a truncated power function of order q is given by

$$\frac{dT_t^q(k)}{dt} = (q-1)(t-k)_+^{q-2} = (q-1) T_t^{(q-1)}(k)$$

then using the definition of divided differences in Section 4.3.2, we have

$$\frac{dM_l(t|q)}{dt} = (-1)^q q (q-1) \left(\frac{[k_{l+1}, \ldots, k_{l+q}] T_t^{(q-1)}(k) - [k_l, \ldots, k_{l+q-1}] T_t^{(q-1)}(k)}{k_{l+q} - k_l} \right)$$

$$= \frac{q}{k_{l+q}-k_l}(-1)^{q-1}(q-1)\left([k_l,\ldots,k_{l+q-1}]T_t^{(q-1)}(k)-[k_{l+1},\ldots,k_{l+q}]T_t^{(q-1)}(k)\right)$$
$$= \frac{q}{k_{l+q}-k_l}\left((-1)^{q-1}(q-1)[k_l,\ldots,k_{l+q-1}]T_t^{(q-1)}(k)-(-1)^{q-1}(q-1)[k_{l+1},\ldots,k_{l+q}]T_t^{(q-1)}(k)\right)$$
$$= \frac{q}{k_{l+q}-k_l}\left(M_l(t|(q-1))-M_{l+1}(t|(q-1))\right).$$

In general, the j^{th} derivative of an M-spline function of order q, $M_l(t|q)$, is

$$\frac{d^jM_l(t|q)}{dt^j}=\frac{q}{k_{l+q}-k_l}\left(\frac{d^{j-1}M_l(t|(q-1))}{dt^{j-1}}-\frac{d^{j-1}M_{l+1}(t|(q-1))}{dt^{j-1}}\right),$$

so the j^{th} derivative of a function which is a linear combination of M-spline basis functions, $f(t)=\sum_{l=1}^{m}\alpha_lM_l(t|q)$, can be given as

$$\frac{d^jf(t)}{dt^j}=\sum_{l=1}^{m}\alpha_l\frac{d^jM_l(t|q)}{dt^j}.$$

7.2.2 Integrals of I-splines

Ramsay (1988) defined the integral of an M-spline of order q as an I-spline which is a piecewise polynomial of order $q+1$. The definition is given in Chapter 4 for a sequence of knots $k_1=k_2=\cdots=k_q<k_{q+1}<\cdots<k_{q+s}<k_{q+s+1}=k_{q+s+2}=\cdots=k_{2q+s}$ used to define an M-spline of order q and $k_h\le t<k_{h+1}$ as

$$I_l(t|q)=\int_a^t M_l(u|q)du,$$

so

$$I_l(t|q)=\begin{cases}0, & l>h\\ \sum_{m=l}^{h}(k_{m+q+1}-k_m)\frac{M_m(t|q+1)}{q+1}, & h-q+1\le l\le h\\ 1 & l<h-q+1,\end{cases}$$

where the lower limit of the integrals is the minimum knot, let it be denoted by a. Based on this definition for the integral of an M-spline we define the integral of an I-spline. Let the integral of $I_l(t|q)$ be denoted by $I_l^1(t|q)$. Using the same sequence of interior knots

employed to define the M-splines, for $k_h \leq t < k_{h+1}$ the integral of an I-spline, $I_l^1(t|q)$, has three different expressions depending on the value of l. For $l > h$ the value of an I-spline is zero so its indefinite integral will be a constant, and hence

$$I_l^1(t|q) = \int_a^t I_l(u|q)du = 0.$$

For $h - q + 1 \leq l \leq h$ an I-spline, $I_l(t|q)$, is given by

$$I_l(t|q) = \sum_{m=l}^{h}(k_{m+q+1} - k_m)\frac{M_m(t|q+1)}{q+1}$$

therefore its integral will be

$$\begin{aligned} I_l^1(t|q) &= \int_a^t \sum_{m=l}^{h}(k_{m+q+1} - k_m)\frac{M_m(u|q+1)}{q+1}du \\ &= \sum_{m=l}^{h}\frac{(k_{m+q+1} - k_m)}{q+1}\int_a^t M_m(u|q+1)du. \end{aligned}$$

$\int_a^t M_m(u|q+1)du$ in the above expression is the integral of an M-spline of order $q+1$ that gives another I-spline, $I_m(t|(q+1)) = \sum_{n=m}^{h}(k_{n+q+2} - k_n)\frac{M_n(t|q+2)}{q+2}$ for $h - q \leq m \leq h$, so

$$I_l^1(t|q) = \sum_{m=l}^{h}\frac{(k_{m+q+1} - k_m)}{q+1}\sum_{n=m}^{h}(k_{n+q+2} - k_n)\frac{M_n(t|q+2)}{q+2}.$$

For $l < h - q + 1$, that is for any value of $t > k_{l+q}$ the value of $I_l(t|q) = 1$. This is because $M_l(t|q) = 0$ for all values of $t > k_{l+q}$. Now the integral of $I_l(t|q)$ has two parts for $t > k_{l+q}$, the integral of the function up to k_{l+q} and from k_{l+q} to t. That is,

$$\int_a^{k_{l+q}} I_l(u|q)du + \int_{k_{l+q}}^t I_l(u|q)du = \left(\sum_{m=l}^{h}\frac{(k_{m+q+1} - k_m)}{q+1}\int_a^{k_{l+q}} M_m(u|q+1)du\right) + (t - k_{l+q}).$$

Therefore, in summary the integral of an I-spline is given by

$$I_l^1(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h}\frac{(k_{m+q+1}-k_m)}{q+1}\sum_{n=m}^{h}(k_{n+q+2} - k_n)\frac{M_n(t|q+2)}{q+2}, & h - q + 1 \leq l \leq h \\ t - k_{l+q} + \sum_{m=l}^{h}\frac{(k_{m+q+1}-k_m)}{q+1}\sum_{n=m}^{h}(k_{n+q+2} - k_n)\frac{M_n(k_{l+q}|q+2)}{q+2}, & l < h - q + 1. \end{cases}$$

The second integral of an I-spline, the integral of $I_l^1(t|q)$ can be obtained in a similar way and is defined as, $I_l^2(t|q) = \int_a^t I_l^1(u|q)du$.

$$I_l^2(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} \int_a^t M_n(u|q+2)du, & h-q+1 \leq l \leq h \\ \frac{t^2}{2} - tk_{l+q} + \frac{k_{l+q}^2}{2} & \\ + \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} \int_a^{k_{l+q}} M_n(u|q+2)du, & l < h-q+1. \end{cases}$$

but $\int_a^t M_n(u|q+2)du$ and $\int_a^{k_{l+q}} M_n(u|q+2)du$ are I-splines of order $q+2$ therefore,

$$I_l^2(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} & \\ \sum_{r=n}^{h} (k_{r+q+3} - k_r) \frac{M_r(t|q+3)}{q+3}, & h-q+1 \leq l \leq h \\ \frac{t^2}{2} - tk_{l+q} + \frac{k_{l+q}^2}{2} + \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} & \\ \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} \sum_{r=n}^{h} (k_{r+q+3} - k_r) \frac{M_r(k_{l+q}|q+3)}{q+3}, & l < h-q+1. \end{cases}$$

Finally the third integral of an I-spline, $I_l^3(t|q) = \int_a^t I_l^2(u|q)du$, is given as

$$I_l^3(t|q) = \begin{cases} 0, & l > h \\ \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} \sum_{r=n}^{h} \frac{(k_{r+q+3}-k_r)}{q+3} & \\ \sum_{v=r}^{h} (k_{v+q+4} - k_v) \frac{M_v(t|q+4)}{q+4}, & h-q+1 \leq l \leq h \\ \frac{t^3}{6} - \frac{t^2 k_{l+q}}{2} + \frac{t k_{l+q}^2}{2} - \frac{k_{l+q}^3}{6} & \\ + \sum_{m=l}^{h} \frac{(k_{m+q+1}-k_m)}{q+1} \sum_{n=m}^{h} \frac{(k_{n+q+2}-k_n)}{q+2} & \\ \sum_{r=n}^{h} \frac{(k_{r+q+3}-k_r)}{q+3} \sum_{v=r}^{h} (k_{v+q+4} - k_v) \frac{M_v(k_{l+q}|q+4)}{q+4}, & l < h-q+1. \end{cases}$$

Now going back to the log-likelihood function, since the exposure-related relative incidence function, $\omega(t-c)$, takes the value 1 in the control periods, $(a_i, c_i]$ and $(d_i, b_i]$, within the observation period, the denominator of the log-likelihood function (7.4) can be

rewritten as

$$\int_{a_i}^{c_i} \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)dt + \int_{c_i}^{d_i} \left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t) \right) \left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t - c_i) \right) dt + \int_{d_i}^{b_i} \sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t)dt$$

Furthermore, the first and the last terms are integrals of only one function, the age-specific relative incidence $\psi(t)$, whereas the second term is the integral of a product of two spline functions. From Chapter 4, we have that the integral of an M-spline of order q is an I-spline of order $q+1$, hence the integral of the linear combination of M-splines can be expressed as a linear combination of I-splines. Therefore, we replace the integrals in the first and third terms by linear combinations of I-spline basis functions which leads to a denominator with the expression

$$\sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(c_i) - \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(a_i) + \int_{c_i}^{d_i} \left(\sum_{l=1}^{m_1} \alpha_l^2 M_{1l}(t) \right) \left(\sum_{l=1}^{m_2} \beta_l^2 M_{2l}(t - c_i) \right) dt + \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(b_i) - \sum_{l=1}^{m_1} \alpha_l^2 I_{1l}(d_i).$$

The $I_{1l}(t)$ are I-splines related to the age effect and $I_{2l}(t)$ will be used to denote I-splines related to the exposure effect. The remaining part in the denominator of the log-likelihood function of the non-parametric SCCS is the nominal risk period $(c_i, d_i]$ where the exposure-related relative incidence can take a value different from 1. This part contains an integral of the product of the two spline functions, $\psi(t)$ and $\omega(t-c)$. In the following subsection we write an expression for this integral.

7.2.3 Integrating the Product of Two Spline Functions

To evaluate the integral of the product of the age-related relative incidence function and exposure-related relative incidence function, $\int_{c_i}^{d_i} \psi(t)\omega(t-c)dt$, we use integration by parts.

Integration by parts makes integrating a product of functions easier by relating them to the integral of their derivative and antiderivative and is defined as follows. Given two continuously differentiable functions $f(t)$ and $g(t)$, the indefinite integral of $f(t)g(t)$ can

be given as

$$\int f(t)g(t)dt = f(t)\int g(t)dt - \int \left(f'(t)\int g(t)dt\right)dt$$

where $f'(t)$ is the first derivative of $f(t)$.

Applying integration by parts to the integral of the product of age and exposure effects in the non-parametric SCCS likelihood, we have

$$\int \psi(t)\omega(t-c)dt = \psi(t)\int \omega(t-c)dt - \int \left(\psi'(t)\int \omega(t-c)dt\right)dt \qquad (7.5)$$

where $\psi'(t)$ is the first derivative of $\psi(t)$. Since $\psi(t)$ and $\omega(t-c)$ are linear combinations of M-spline basis functions, $\int \omega(t-c)dt$ can be expressed as a linear combination of I-splines denoted by $I_E(t-c)$

$$I_E(t-c) = \int_c^t \omega(u-c)du = \int_c^t \sum_{l=1}^{m_2} \beta_l^2 M_{2l}(u-c)du = \sum_{l=1}^{m_2} \beta_l^2 I_{2l}(t-c).$$

Letting the integral of the linear combination of I-splines, $I_E(t-c)$ be denoted by $I_E^1(t-c)$, the integral of $I_E^1(t-c)$ by $I_E^2(t-c)$ and the integral of $I_E^2(t-c)$ by $I_E^3(t-c)$,

$$I_E^1(t-c) = \int I_E(t-c)dt, \quad I_E^2(t-c) = \int I_E^1(t-c)dt \quad \text{and} \quad I_E^3(t-c) = \int I_E^2(t-c)dt,$$

so the expression in Equation (7.5) becomes

$$\int \psi(t)\omega(t-c)dt = \psi(t)I_E(t-c) - \int (\psi'(t)I_E(t-c))\,dt.$$

The last term of this equation is again an integral of a product of two non-constant functions. We therefore apply integration by parts repeatedly until none of the terms is an integral of two non-constant functions:

$$\int \psi(t)\omega(t-c)dt = \psi(t)I_E(t-c) - \int (\psi'(t)I_E(t-c))\,dt$$

$$
\begin{aligned}
&= \psi(t)I_E(t-c) - \left[\psi'(t)\int I_E(t-c)dt - \int\left(\psi''(t)\int I_E(t-c)dt\right)dt\right] \\
&= \psi(t)I_E(t-c) - \psi'(t)I_E^1(t-c) + \int \psi''(t)I_E^1(t-c)dt \\
&= \psi(t)I_E(t-c) - \psi'(t)I_E^1(t-c) + \psi''(t)\int I_E^1(t-c)dt - \int\left(\psi'''(t)\int I_E^1(t-c)dt\right)dt \\
&= \psi(t)I_E(t-c) - \psi'(t)I_E^1(t-c) + \psi''(t)I_E^2(t-c) - \int\left(\psi'''(t)I_E^2(t-c)dt\right)dt \\
&= \psi(t)I_E(t-c) - \psi'(t)I_E^1(t-c) + \psi''(t)I_E^2(t-c) - \psi'''(t)\int I_E^2(t-c)dt \\
&= \psi(t)I_E(t-c) - \psi'(t)I_E^1(t-c) + \psi''(t)I_E^2(t-c) - \psi'''(t)I_E^3(t-c)
\end{aligned}
$$

where $\psi'(t)$, $\psi''(t)$ and $\psi'''(t)$ are the first, second and third derivatives of $\psi(t)$ respectively. $\psi'''(t)$ is a constant function that does not depend on t because $\psi(t)$ is a piecewise cubic function. Therefore, the integral of the product of $\psi(t)$ and $\omega(t-c)$ in the nominal risk period $(c_i, d_i]$ is

$$
\begin{aligned}
\int_{c_i}^{d_i}\psi(t)\omega(t-c_i)dt &= \left(\psi(d_i)I_E(d_i-c_i) - \psi'(d_i)I_E^1(d_i-c_i) + \psi''(d_i)I_E^2(d_i-c_i) - \psi'''(d_i)I_E^3(d_i-c_i)\right) \\
&\quad - \left(\psi(c_i)I_E(0) - \psi'(c_i)I_E^1(0) + \psi''(c_i)I_E^2(0) - \psi'''(c_i)I_E^3(0)\right).
\end{aligned}
$$

Then the log-likelihood function of the non-parametric SCCS method, obtained by replacing the appropriate expressions for the $\int_{a_i}^{c_i}\psi(t)dt$, $\int_{c_i}^{d_i}\psi(t)\omega(t-c)dt$ and $\int_{d_i}^{b_i}\psi(t)dt$ in the denominator, is

$$
l = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\left(\frac{\left(\sum_{l=1}^{m_1}\alpha_l^2 M_{1l}(t_{ij})\right)\left(\sum_{l=1}^{m_2}\beta_l^2 M_{2l}(t_{ij}-c_i)\right)^{I(c_i<t_{ij}\le d_i)}}{B}\right) \tag{7.6}
$$

where

$$
\begin{aligned}
B &= \sum_{l=1}^{m_1}\alpha_l^2 I_l(c_i) - \sum_{l=1}^{m_1}\alpha_l^2 I_l(a_i) + \sum_{l=1}^{m_1}\alpha_l^2 I_l(b_i) - \sum_{l=1}^{m_1}\alpha_l^2 I_l(d_i) \\
&\quad + \left(\psi(d_i)I_E(d_i-c_i) - \psi'(d_i)I_E^1(d_i-c_i) + \psi''(c_i)I_E^2(d_i-c_i) - \psi'''(d_i)I_E^3(d_i-c_i)\right) \\
&\quad - \left(\psi(c_i)I_E(0) - \psi'(c_i)I_E^1(0) + \psi''(c_i)I_E^2(0) - \psi'''(d_i)I_E^3(0)\right)
\end{aligned}
$$

and $I_E^1(t-c) = \sum_{l=1}^{m_2}\beta_l^2 I_{2l}^1(t-c)$, $I_E^2(t-c) = \sum_{l=1}^{m_2}\beta_l^2 I_{2l}^2(t-c)$, $I_E^3(t-c) = \sum_{l=1}^{m_2}\beta_l^2 I_{2l}^3(t-c)$

$I^1_{2l}(t-c)$, $I^3_{2l}(t-c)$ and $I^3_{2l}(t-c)$ are the first, second and third integrals of the l^{th} I-spline ($I_{2l}(t-c)$) related to exposure, respectively.

7.3 Penalised Log-Likelihood

The numbers of knots, which determine the numbers of M-spline basis functions that make up the age-specific and exposure-related relative incidence functions are chosen a priori. Maximising the log-likelihood function (7.6) after choosing too large a number of knots over-fits the true curves, while selecting too small a number of knots leads to under-fitting overly smoothed curves. Therefore, to control the smoothness of the estimated functions we fix the numbers of knots at higher values than are believed to be enough to fit the functions and introduce roughness penalty terms to the log-likelihood function (7.6). Following Joly and Commenges (1999), we choose a roughness measure to be the sum of the square norms of the second derivatives of the age and exposure effect functions. This leads to the penalised log-likelihood function

$$\begin{aligned} pl &= l(\boldsymbol{\alpha},\boldsymbol{\beta}) - \lambda_1 \int \left(\sum_{l=1}^{m_1} \alpha_l^2 M''_{1l}(u)\right)^2 du - \lambda_2 \int \left(\sum_{l=1}^{m_2} \beta_l^2 M''_{2l}(u)\right)^2 du \\ &= l(\boldsymbol{\alpha},\boldsymbol{\beta}) - \lambda_1((\boldsymbol{\alpha^2})^T \mathbf{A}_1 \boldsymbol{\alpha^2}) - \lambda_2((\boldsymbol{\beta^2})^T \mathbf{A}_2 \boldsymbol{\beta^2}) \end{aligned} \tag{7.7}$$

where $\boldsymbol{\alpha}$ is a vector of parameters $\alpha_1,\ldots,\alpha_{m_1}$, that define the age-specific relative incidence function and $\boldsymbol{\alpha^2} = \alpha_1^2,\ldots,\alpha_{m_1}^2$, $\boldsymbol{\beta^2} = \beta_1^2,\ldots,\beta_{m_2}^2$ are parameters related to the exposure effect, $\mathbf{A}_1$ is an $m_1 \times m_1$ matrix with (r,l) element $\int M''_{1r}(u)M''_{1l}(u)du$, $\mathbf{A}_2$ is an $m_2 \times m_2$ matrix with (r,l) element $\int M''_{2r}(u)M''_{2l}(u)du$, $l(\boldsymbol{\alpha},\boldsymbol{\beta})$ is the log-likelihood function (7.6). λ_1 and λ_2 are non-negative smoothing parameters that control the trade off between the model fit and smoothness of the functions. So the penalised log-likelihood function (7.7) is maximised, for fixed λ_1 and λ_2 values, to estimate the parameters related

to age and exposure effects.

We choose the smoothing parameters using the approximated cross validation method discussed in Chapters 5 and 6. λ_1 is first chosen by maximising the cross validation score presented in Chapter 5 by taking no exposure effect then λ_2 is chosen by maximising the cross validation score in Chapter 6 taking the age effect to be zero. The scores to be maximised to select λ_1 and λ_2 respectively are:

$$\bar{V}_1(\lambda_1) = l(\hat{\boldsymbol{\alpha}}) - \text{tr}([\hat{H}_1 - 2\lambda_1 \mathbf{S}_1]^{-1}\hat{H}_1), \tag{7.8}$$

and

$$\bar{V}_2(\lambda_2) = l(\hat{\boldsymbol{\beta}}) - \text{tr}([\hat{H}_2 - 2\lambda_2 \mathbf{S}_2]^{-1}\hat{H}_2), \tag{7.9}$$

where $l(\hat{\boldsymbol{\alpha}})$ is the log-likelihood function in Equation (7.6) where exposure effect is taken to be zero and evaluated at the maximum penalised likelihood estimates ($\hat{\boldsymbol{\alpha}}$). $\hat{H}_1 = \frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}(\hat{\boldsymbol{\alpha}})$ is the log-likelihood part of the Hessian of the penalised log-likelihood, taking zero exposure effect, evaluated at the penalised maximum likelihood estimates $\hat{\boldsymbol{\alpha}}$. $\mathbf{S}_1 = 4\left(\mathbf{A}_1 o(\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\right) + 2(\text{diag}(\mathbf{A}_1\boldsymbol{\alpha}^2))$. Similarly, $l(\hat{\boldsymbol{\beta}})$ is the log-likelihood (7.6) taking age effect to be zero, $\hat{H}_2 = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}})$ is the Hessian when the age effect is considered to be zero and $\mathbf{S}_2 = 4\left(\mathbf{A}_2 o(\boldsymbol{\beta}\boldsymbol{\beta}^T)\right) + 2(\text{diag}(\mathbf{A}_2\boldsymbol{\beta}^2))$.

In Chapter 5, we showed that the parameter of interest, the exposure-related relative incidence value, is not unduly sensitive to changes in the smoothing parameter that controls the roughness of age-specific relative incidence function. Therefore, in this chapter an alternative approach is to consider the smoothing parameter related to age effect, λ_1, to be fixed at some reasonable value. Then after choosing the smoothing parameters the log-likelihood function (7.7) is maximised for fixed λ_1 and λ_2.

7.4 Simulation Study

To evaluate the performance of the new non-parametric SCCS method and to compare it with the extensions made to the standard SCCS method in Chapters 5 and 6, we conducted a simulation study.

7.4.1 Design of the Simulation Study

The number of cases used in this simulation was 1000, each with ages at the start and end of the observation period of 0 and 730 days respectively. For each case, the risk period between the start of exposure c_i and end of exposure d_i was taken as 49 days. The baseline incidence was generated from a sine function, defined as $\lambda_0(t) \propto 8(sin(0.01 \times t)) + 9$ at age t. The true age-related relative incidence function is presented in Panel a of Figure 7.1. Ages at start of exposure c_i, for $i : 1, \ldots, 1000$, were sampled within (0,730] from an exponential density with rate 0.003. The histogram of c_i is shown in Panel b of Figure 7.1.

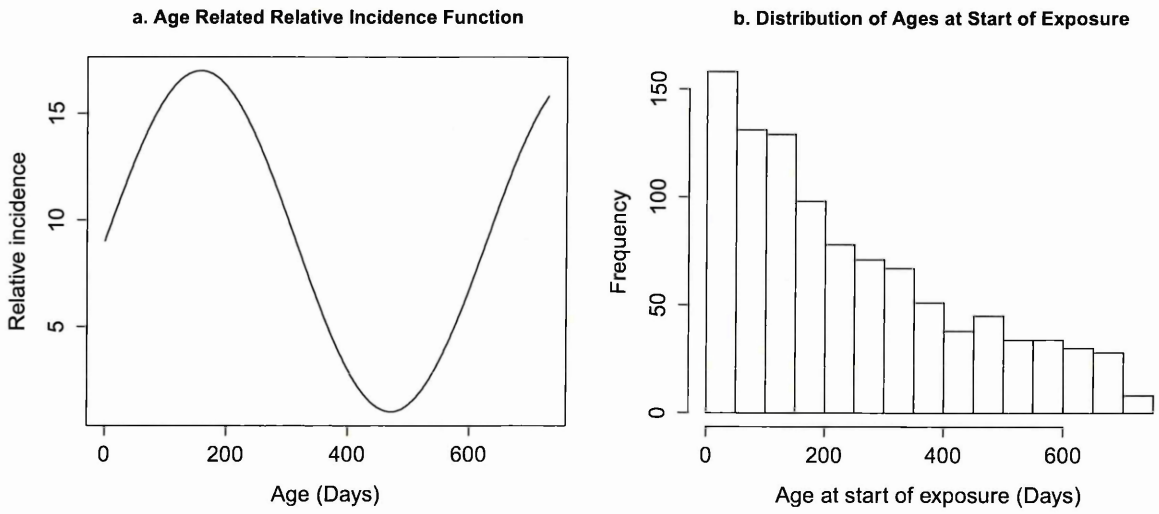

Figure 7.1: *True age-related relative incidence function in Panel (a) and distribution of ages at start of exposure in Panel (b), which were used to simulate data sets*

For the given age-related relative incidence function and distribution of age at exposure, we investigated four scenarios of exposure-related relative incidence function, $\omega(t-c)$. These functions take a value one outside the risk period $(c_i, d_i]$, that is when time since start of exposure $t - c \leq 0$ or $t - c > 49$. These scenarios were also used in Chapter 6 and are presented in Figure 6.1.

Without loss of generality we consider each case to have experienced only one event. Then the daily incidence rates within the observation period are evaluated as the product of the age-related relative incidence and the exposure-related relative incidence. An event day for each individual was generated from a multinomial distribution. The probability of an event at a given day within the observation period was computed as the incidence rate for that day divided by the sum of the rates for all the days within the observation period. For each scenario 100 data sets were simulated.

7.4.2 Analysis

The data sets generated were analyzed by the three new versions of SCCS presented in this thesis: (1) smooth age effect with parametric exposure effect (step function) (2) parametric age effect (step function) with spline-based exposure effect and (3) the non-parametric SCCS proposed in this chapter.

For the first method, seven exposure groups of length seven days between 0 and 49 were chosen to represent the exposure effect by a step function. For methods (1) and (3), to represent the age effect with a spline function 9 interior knots between the minimum of ages at the start of observation (zero) and the maximum of the ages at the end of observation periods (730) were chosen. For the age effect, since exposure-related parameters are not duly sensitive to changes in the smoothing parameter related to age effect, we chose

a smoothing parameter for the first sample in a given scenario by the cross validation method and used the same value for the remaining samples.

For the second method, where age is represented with a piecewise constant function, six age groups with cut points at 0, 120, 240, 360, 480, 600 and 730 days were pre-specified. To represent the exposure effect with a spline function in methods (2) and (3), a nominal risk period of 49 days was chosen. 12 interior knots between zero and 49 were selected. The smoothing parameter of the exposure was chosen by the cross validation method for all the samples in the two methods. In addition, we fitted method (2), but with only three age groups with cut points at 0, 240, 480 and 730 days, to see how a change in age groups affects the results.

To compare the performance of the three methods we used the mean of integrated squared errors (MISE) (see Sections 5.3 and 6.2 for the definition of MISE) and their standard deviations (SD) in estimating the age and exposure-related relative incidence functions. To compute the MISE and SD related to the age effect, we used the cumulative age-specific relative incidence function, for the true and estimated functions, constrained to have a maximum value of one to make the three methods comparable.

7.4.3 Results

Results of the simulation study are presented in this section. Table 7.1 presents the MISE and SD results in estimating the age and exposure effects using the three methods developed in this thesis. The method proposed in Chapter 6 was fitted twice for each generated data set using 6 and 3 age groups.

Table 7.1: *Mean integrated squared error (MISE) and standard deviation (SD) obtained from the three spline-based SCCS methods: SCCS with smooth age effect, SCCS with smooth exposure effect (twice with 6 and 3 age groups) and SCCS with both age and exposure effects represented by splines. Each simulated data set was fitted by the three methods using a nominal risk period of 49 days. The true age-specific relative incidence function was generated from sine function*

	Smooth age	Smooth exposure 6 age groups	Smooth exposure 3 age groups	Smooth age & exposure
		Scenario 1		
Effects	MISE (SD)	MISE (SD)	MISE (SD)	MISE (SD)
Exposure	13.182 (6.581)	7.318 (4.792)	7.393 (4.835)	7.220 (4.433)
Age	0.110 (0.103)	0.181 (0.086)	1.466 (0.102)	0.110 (0.106)
		Scenario 2		
Exposure	22.959 (10.249)	10.849 (12.996)	10.507 (12.678)	9.298 (7.188)
Age	0.117 (0.105)	0.202 (0.107)	1.483 (0.102)	0.123 (0.106)
		Scenario 3		
Exposure	9.856 (5.597)	5.438 (6.466)	5.552 (6.597)	4.393 (4.372)
Age	0.107 (0.089)	0.187 (0.093)	1.476 (0.111)	0.109 (0.090)
		Scenario 4		
Exposure	10.007 (4.882)	6.388 (8.451)	6.424 (8.207)	4.890 (6.328)
Age	0.126 (0.108)	0.204 (0.103)	1.490 (0.121)	0.129 (0.107)

The results in Table 7.1 suggest that the new method performs well. In estimating the age-specific relative incidence function the non-parametric method has equivalent performance as method (1) with smooth age effect and has better performance as compared to method (2).

In estimating the exposure-related relative incidence function, the non-parametric method showed the highest performance as compared to both methods (1) and (2). For method (2), when the age groups used in modelling the age effect are reduced to three, the performance of the method reduces, which indicates that mis-specification of age groups

may lead to a reduced performance of this method. However, for scenario2 surprisingly the performance increased when the number of age groups is reduced. The non-parametric method developed in this chapter does not have a limitation related to mis-specification of age and exposure groups.

The estimated age-related and exposure-related relative incidence functions along with their true curves are presented in Figures 7.2, 7.3, 7.4 and 7.5 for scenarios 1, 2, 3 and 4 respectively (the model with three age groups is not presented). The curves related to the age effect are plotted by constraining the cumulative relative incidence at the maximum of the ages at the end of observation period to be one.

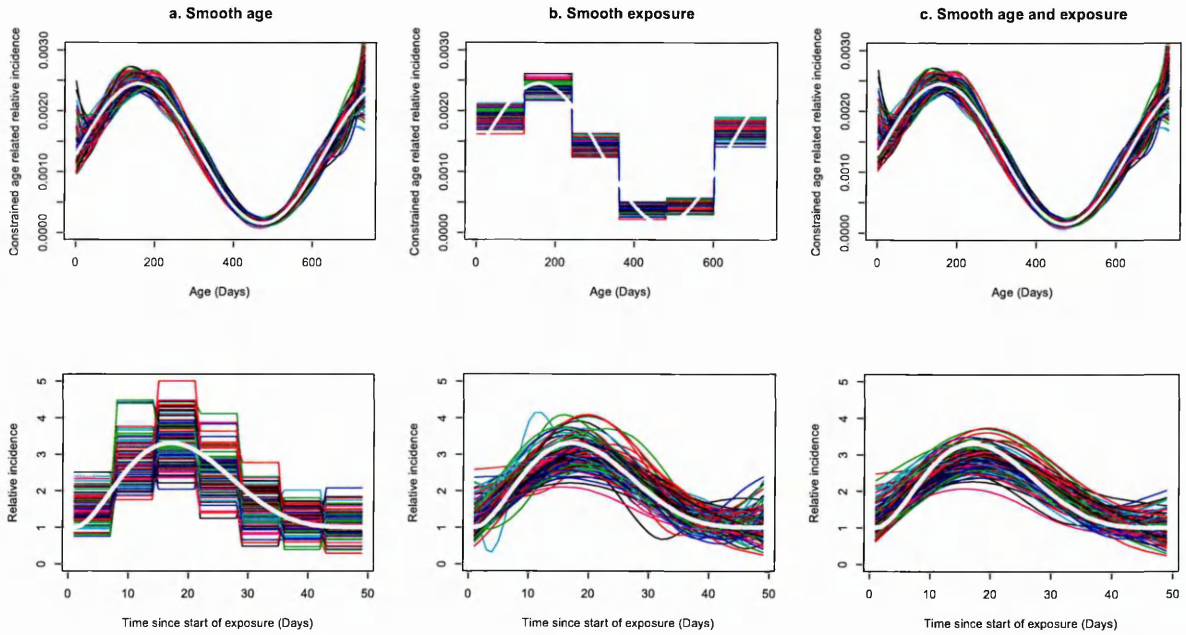

Figure 7.2: *Estimated relative incidence curves for scenario* 1*; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

The figures suggest that the non-parametric method seems to perform well in estimating both the age and exposure-related relative incidence curves. In all the cases the true functions are within the range of the estimated curves and the estimated curves seem to follow the trend of the true functions. However there are some estimated exposure-related curves that over-fitted the true curve for scenario 2, (Figure 7.3), where the true function is a constant. These could be due to numerical problems in choosing the smoothing parameter.

The performance of the three methods is reduced for scenario 2 where the true exposure-related relative incidence function is a constant function.

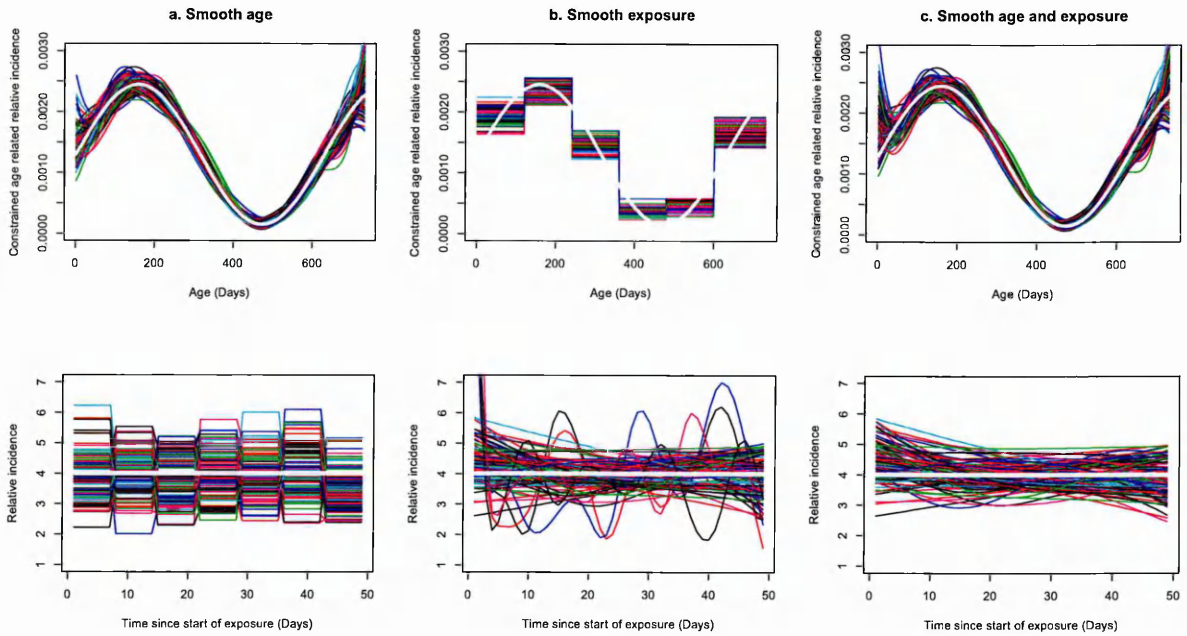

Figure 7.3: *Estimated relative incidence curves for scenario 2; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

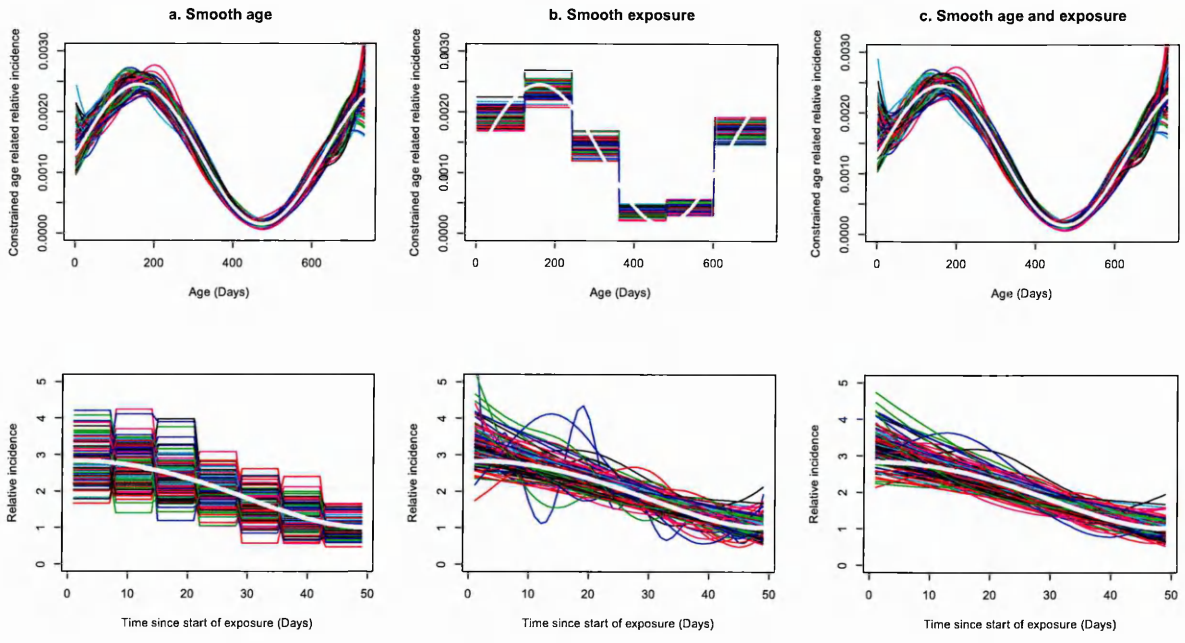

Figure 7.4: *Estimated relative incidence curves for scenario 3; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

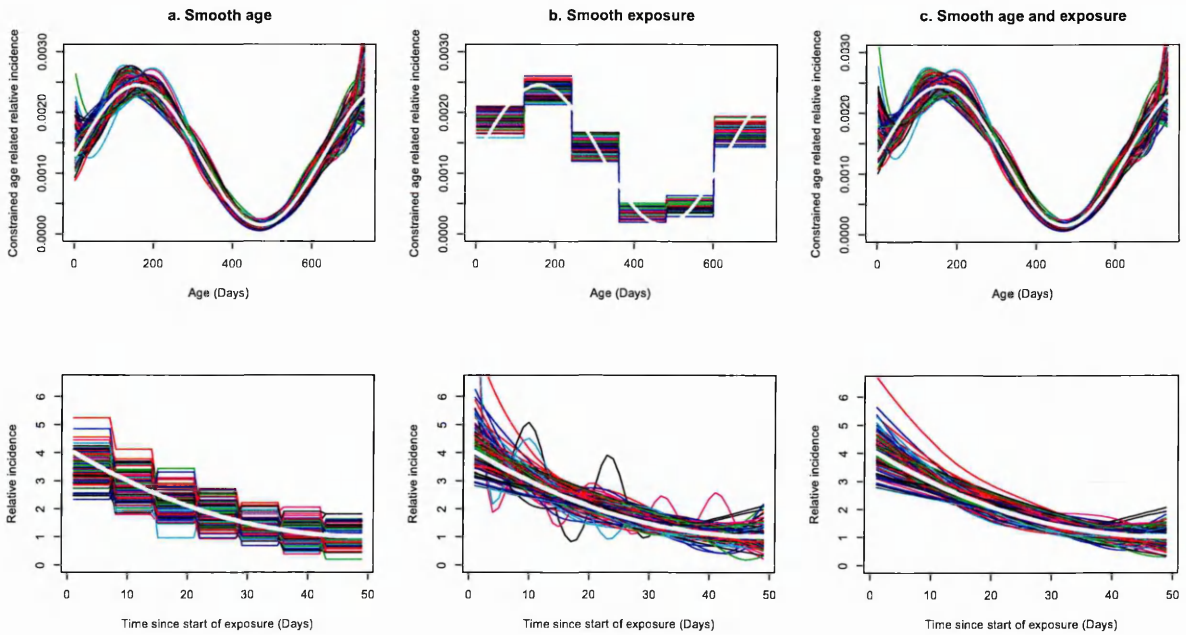

Figure 7.5: *Estimated relative incidence curves for scenario 4; the top panels show age-related relative incidence curves and the bottom panels exposure-related relative incidence curves. In panels a are results from SCCS with smooth age effect, panels b SCCS with smooth exposure effect and panels c SCCS with both age and exposure represented with splines. The white solid lines in all panels represent the true functions.*

7.5 Application

We illustrate the non-parametric self-controlled case series method by applying it to data on MMR vaccines and febrile convulsions. The data were introduced and described in Chapter 5. The number of cases in the data set is 2,389 children aged between 29 and 730 days with 3,826 events. As in Chapter 6, we chose the nominal risk period post MMR vaccine to be 50 days.

Linear combinations of cubic M-splines are used to represent the age and exposure effects. For the MMR vaccine related relative incidence function we used 12 equally

spaced knots between 0 and 50. The smoothing parameter λ_2 for the exposure effect was chosen by the cross validation method and was found to be 0.031. For the age-related relative incidence, we used 12 interior knots and chose the smoothing parameter using the cross validation method by keeping the exposure effect zero. The value selected was 1.07×10^9. Then for the given values of the smoothing parameters, we maximised the non-parametric SCCS penalised log-likelihood function (7.7). The estimated age and exposure-related relative incidence curves are presented in Figure 7.6.

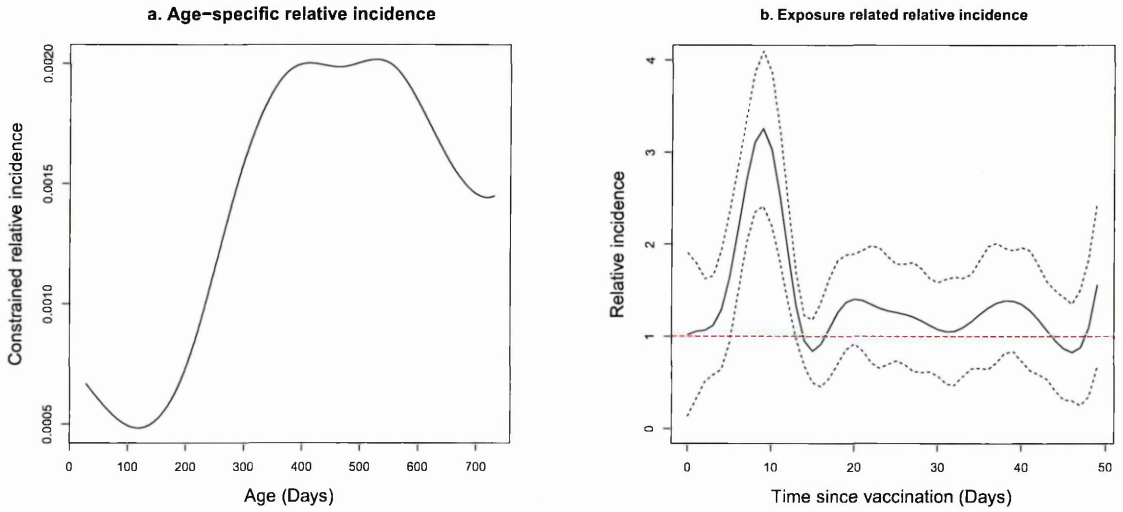

Figure 7.6: *Relative incidence curves estimated by fitting non-parametric SCCS. Panel (a) shows the estimated constrained age-related relative incidence function Panel (b) represents estimated exposure-related relative incidence curve (solid line) along with 95% confidence bands denoted by the dashed lines*

Panel (a) of Figure 7.6 shows the estimated age-related relative incidence function, where the cumulative age effect is constrained to have a value one at the maximum end of observation period. This figure is similar to the curve in Figure 5.7 in which the age effect was estimated based on splines but the exposure effect was estimated by a step function. Panel (b) of the figure shows the relative incidence curve post MMR vaccine. From the

figure, it can be seen that there is a significant increase in the risk of febrile convulsion from six to 12 days after exposure to MMR vaccine. Five and 13 days after vaccination have a borderline insignificant risk of febrile convulsion. There is no increased risk in other periods. These results are similar to the results obtained from the MMR exposure-related relative incidence function estimated in Section 6.3 of Chapter 6. However there is a slight difference in the results, in Figure 6.6 there is an increased risk of febrile convulsion between 19 and 21 days following an MMR vaccine but not in panel (b) of Figure 7.6. This difference could be because of modelling age effect using a step function in Chapter 6.

The confidence bands for the exposure-related relative incidence function were evaluated using the approximate method presented in Section 6.1.1 of Chapter 6. However, the 95% coverage probabilities of these confidence bands in the current setting need to be studied. An alternative method to use is bootstrap method as suggested by Joly and Commenges (1999).

7.6 Discussion

The extension developed here combines the extensions developed in Chapters 5 and 6. In Chapter 5, only the age effect was approximated by a linear combination of M-spline basis functions and the exposure effect was represented by a piecewise constant function. In Chapter 6, splines were used only to estimate the exposure-related relative incidence function and age was taken into account based on step functions. In this chapter, the effects of both age and exposure in the SCCS model are represented by linear combinations of M-spline basis functions simultaneously. The new method avoids the mis-specification bias that may occur due to poor choice of age or exposure groups in the previous two chapters due to the use of step functions.

The denominator of the log-likelihood function of the new method includes the integral of a product of two spline functions, namely the age-related and the exposure-related relative incidence functions. Rather than using numerical integration techniques, we evaluated this integral analytically using integration by parts. This required evaluation of the first, second and third integrals of an I-spline function, based on the definition of the integral of an M-spline given by Ramsay (1988).

A simulation study was conducted to evaluate the performance of the new method, non-parametric SCCS. It was found that the new method has good performance as compared to the extensions presented in Chapters 5 and 6. According to the results the new method has comparable or better performance to the previous two extensions. The new method also has an advantage over the others in the fact that the other methods can give biased estimates if the a priori specified age or exposure groups are poorly or mis-specified. In addition, the non-parametric method avoids the limitation of the extension in Chapter 6 that if the age groups used to represent the age-specific relative incidence are too many, the method may go into computational problems.

Chapter 8

General Conclusions and Further Research

The self-controlled case series method is one of the study designs that are used to investigate safety of vaccines and other drugs after they are licensed for use. The main advantages of the SCCS method over cohort and case-control methods are that it is only based on information on individuals with a disease event (cases), so no separate controls are required, it implicitly controls all fixed confounding variables and under many circumstances it has good efficiency (Whitaker *et al.*, 2006).

8.1 Summary and Conclusions

The SCCS method does not automatically control for time-varying confounders, so they have to be identified and included in the model. There are two ways of including the time-varying covariate age: the standard (parametric) SCCS that uses step functions by specifying age groups a priori and the semi-parametric SCCS which leaves the shape of the age effect unspecified. A review of SCCS studies on safety of vaccines since 1995

when the method was first published up to the beginning of 2014, showed that in general the method was applied appropriately. Of the reviewed studies, 81 used the standard SCCS method, only one used the semi-parametric method, one reported that age was included as a continuous variable and one study used linear and quadratic functions to model the age effect. One study used fractional polynomials to estimate the steps of a piecewise constant age effect. Four of the reviewed studies used an extended version of the standard SCCS method that allows event dependent future exposures.

In Chapter 2, a simulation study was conducted to investigate the limitations of the parametric and the semi-parametric SCCS methods. The results showed that in the parametric SCCS the age groups which should be chosen a priori may lead to biased exposure-related relative incidence estimates if they are incorrectly specified. On the other hand, fitting the semi-parametric method may become impossible to compute (at least in R) when the number of cases is moderately large. For example, the semi-parametric method does not run for more than 500 cases using the R software package.

To circumvent these limitations, we proposed modelling the effect of age in the SCCS method using smooth functions, namely a linear combination of cubic M-splines (Ghebremichael-Weldeselassie *et al.*, 2014a). Spline methods are curve fitting methods which have a flavour of both parametric and non-parametric methods. They are piecewise polynomials (and hence parametric functions) connected at points known as knots. Regression splines and smoothing splines are types of splines which differ depending on the number of knots used to connect the pieces of polynomials. Smoothing splines take data points to be the knots, hence a large number of parameters may need to be estimated; regression splines use fewer knots. Penalised regression splines are a compromise between regression and smoothing splines. This is the approach we used in the thesis. Splines are more flexible than global

parametric functions and give biologically plausible shapes. We used penalised regression splines based on M-splines. In the SCCS model age effects should be non-negative since they are relative effects. M-splines are particularly useful in the SCCS model because they are positive functions and their linear combination can be non-negative by constraining the coefficients to be non-negative. In addition, the integral in the SCCS likelihood function can be obtained analytically by the use of M-splines because their integrals are I-splines.

A simulation study showed that the new method developed in Chapter 5 has a higher or equivalent performance to the semi-parametric and standard SCCS methods with correctly specified age groups when the sample size is moderate. The new method showed an improved performance as compared to the standard SCCS with mis-specified age groups. Moreover, unlike the semi-parametric method, the new spline-based method works well for large data sets.

Estimation of parameters in this method is based on a penalised log-likelihood function where the smoothing parameter attached to the penalty term is chosen by using a cross validation method. It was found that the parameters related to the exposure effect are not overly sensitive to changes in the smoothing parameter value.

In Chapter 6 we proposed using a linear combination of cubic M-splines to represent the exposure effect (time since start of exposure) to avoid the limitations of using a step function (Ghebremichael-Weldeselassie *et al.*, 2014b). Similar to the first extension developed, this method also showed an increased performance over the standard SCCS method based on the simulation studies conducted. This new method is particulary useful when the risk period is long. For example to investigate the association between oral antibiotic prescription and pregnancy (Petersen *et al.*, 2010), where the event outcome is prescrip-

tion of oral antibiotic and exposure the nine months of pregnancy. Another example is to investigate the adverse effect of a point exposure to idiopathic thrombocytopenic purpura (ITP) vaccine which has a risk period of 42 days (Miller *et al.*, 2001). The method is also useful when the the risk period is unbounded. The risk period is said to be unbounded when the risk period of the cases ends at the end of observation. For example, to investigate the association between exposure to thiazolidinedione use which could last for several years and an outcome event of fracture (Douglas *et al.*, 2009). The method can be applied when the cases have identical or varying risk lengths.

And finally an extension that combines the first two extensions was developed in Chapter 7 where age and exposure effects were modelled by spline functions. We used a linear combination of M-splines. In order to fit both effects with flexible functions at the same time, we developed first, second and third integrals of an I-spline. This method was evaluated by a simulation study that showed a good performance. The non-parametric SCCS method does not suffer from mis-specification bias unlike the first two extensions of the standard and the semi-parametric SCCS methods.

The method proposed in Chapter 5 is important when the adverse outcome varies widely with respect to age over the observation period, which may be particularly true in child and elderly populations. It may also be useful to allow for strong seasonal effects when the underlying time line is calendar time. This method could also prove useful when no prior knowledge of appropriate age effects is available. The methods in Chapters 6 and 7 will be most useful when there is no prior hypothesis about the risk period, or the way in which risk changes over the risk period is of interest. They can be used to determine appropriate exposure groups to be used with the standard SCCS method then the exposure-related relative incidences obtained from both the standard and spline-based

methods can be plotted and compared.

8.2 Future Research

The method developed in Chapter 5, where age was included as a spline function, was applied to investigate the association between paediatric vaccines and febrile convulsions. It was observed that the use of splines in place of a step function to represent the age effect resulted in a notable difference in the relative incidence of exposure to DTP vaccine. This result shows that mis-specification of the age effect might result in significant bias in the exposure-related relative incidence function. From the review in Chapter 2, several studies including Ali *et al.* (2005); Burwen *et al.* (2006); Juurlink *et al.* (2006); Zinman *et al.* (2009) excluded age effects from their analyses since their observation periods were short. In this respect a simulation study to investigate the effect of ignoring the age effect on exposure parameters when observation periods are short may be useful.

Further extension of the spline-based SCCS method developed in Chapter 6 to non-vaccine pharmacoepidemiology, notably to incorporate the effect of dose within a more general weighted cumulative exposure model framework, would be desirable. Moreover, further extension, in terms of incorporating more than one exposure at the same time to assess their association to a single outcome event, would be useful. However, if the exposures do not overlap it may be possible to use the developed approach; overlapping exposures would lead to a product of two spline functions (related to the two exposures) within the overlapping intervals. With no overlaps, the relative incidence at a given point in an interval is the product of the age-related relative incidence and a relative incidence related to one of the two exposures. So a second exposure can be included in the log-likelihood function (6.4) by multiplying the numerator of the function by a

linear combination of cubic M-splines for the second exposure with an indicator variable similar to the first one. And in the denominator the exponent of e_{lh} in the function is multiplied by an indicator for the second exposure $(1 - I(s_i \leq l_{ih} < f_i))$, where s_i and f_i are the ages at start and end of the exposure respectively, and an expression similar to the first exposure multiplies the denominator. If the two exposures overlap, there will be a product of two spline functions (related to the two exposures) in the log-likelihood function, therefore a similar approach to the method developed in Chapter 7 can be used, while the age effect is represented by a step function.

Another extension to the method developed in Chapter 6 is to include a washout period effect, which would be straightforward to include as a step function. A washout period could also easily be included as a spline function if exposure periods are all of the same length. If exposure periods are of differing lengths this would be more difficult because each individual's exposure period will end at a different level.

The performance (coverage probabilities) of the approximate confidence bands used in Section 7.5 of Chapter 7 could be evaluated further by simulations.

The spline-based methodologies developed in this thesis require the assumptions stated in Section 2.1.1 of Chapter 2 to be met. However, the SCCS method has been extended in order to weaken the assumptions required. Farrington *et al.* (2009) extended a method to allow non-exogenous exposures and Kuhnert *et al.* (2011) developed a method to handle event-dependent exposures and deaths. These other extensions allow event dependent observation periods (Farrington *et al.*, 2011) and dependent recurrences (Farrington and Hocine, 2010). Simpson (2013) extended the standard SCCS to allow the occurrence of an event to increase the future event risk. To this end, spline-based methods that incorporate these extensions may be useful.

The estimation methods in all the extensions developed in this thesis involve a two step procedure, selecting the smoothing parameter of one variable taking the other from the log-likelihood out then estimate all the required parameters for a fixed value of the smoothing parameter. It might therefore, be worthwhile to explore methods that estimate parameters in a single step.

Post-licensure studies of vaccines and other drugs are often conducted to investigate their safety against rare events and since the SCCS method uses only cases (individuals who experienced the event), such studies may only have small numbers of cases available. Fitting piecewise cubic polynomials (spline functions) to small data sets could be difficult, but kernel smoothers can be fitted even for a small number of observations. Therefore, the use of kernel smoothers in the SCCS context for small sample sizes, and of course for large sample sizes should they offer any improvements over splines, may be worthwhile to investigate.

The review of vaccine studies in Chapter 3 showed that only small number of studies applied the extensions of the standard SCCS method. This may be because the extensions are much more technically challenging than the basic SCCS model. Therefore providing accessible software tools to implement these extensions in a unified framework within the standard software packages and preparation of tutorials is important.

8.3 Final Remarks

The methodologies developed in this thesis greatly improve the performance of the self-controlled case series method in estimating both the age effect and time-varying exposure effects. They avoid the limitations of the parametric and semi-parametric SCCS methods. The sensitivity of the parametric SCCS to mis-specification of age groups is avoided by the

extensions developed in Chapters 5 and 7. In estimating time-varying exposures, there is no need to pre-specify exposure groups in the methodologies developed in Chapters 6 and 7 unlike the parametric and semi-parametric SCCS methods. All the methods developed can be applied to data sets with large number of cases.

Bibliography

Abrahamowicz, M., Bartlett, G., Tamblyn, R., and du Berger, R. (2006). Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *Journal of Clinical Epidemiology* **59(4)**, 393–403.

Ali, M., Do, C. G., Clemens, J. D., et al. (2005). The use of a computerized database to monitor vaccine safety in Viet Nam. *Bulletin of the World Health Organization* **83**, 604–610.

Altman, D. G. (1991). Categorizing continuous variables. *British Journal of Cancer* **64(5)**, 975–975.

Andrews, N. J. (2002). Statistical assessment of the association between vaccination and rare adverse events post-licensure. *Vaccine* **20**, S49–S53.

Andrews, N., Miller, E., Waight, P., Farrington, P., Crowcroft, N., Stowe, J., and Taylor, B. (2001). Does oral polio vaccine cause intussusception in infants? Evidence from a sequence of three self-controlled cases series studies in the united kingdom. *European Journal of Epidemiology* **17**, 701–706.

Andrews, N., Miller, E., Taylor, B., Lingam, R., Simmons, A., Stowe, J., and Waight, P. (2002). Recall bias, MMR, and autism. *Archives of Disease in Childhood* **87**, 493–494.

Andrews, N., Stowe, J., Miller, E., and Taylor, B. (2007). Post-licensure safety of the meningococcal group C conjugate vaccine. *Human Vaccines* **3**, 59–63.

Andrews, N., Stowe, J., Wise, L., and Miller, E. (2010). Postlicensure comparison of the safety profile of diphtheria/tetanus/whole cell pertussis/haemophilus influenza type b vaccine and a 5-in-1 diphtheria/tetanus/acellular pertussis/haemophilus influenza type b/polio vaccine in the United Kingdom. *Vaccine* **28**, 7215–7220.

Barlow, W.E., Davis, R.L., Glasser, J.W., Rhodes, P.H., Thompson, R.S., Mullooly, J.P., Black, S.B., Shinefield, H.R., Ward. J.I., and Marcy, S.M. (2001). The risk of seizures after receipt of whole-cell pertussis or measles, mumps, and rubella vaccine. *New England Journal of Medicine* **345,** 656–661.

Becker, N. G., Li, Z., Kelman, C.W. (2004). The effect of transient exposures on the risk of an acute illness with low hazard rate. *Biostatistics* **5**, 239-248.

Berhane, K., Hauptmann, M., and Langholz, B. (2008). Using tensor product splines in modeling exposure-time-response relationships: application to the Colorado Plateau uranium miners cohort. *Statistics in Medicine* **27,** 5484–5496.

Breslow, N.E., Lubin, J.H., Marek, P., and Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Society* **78(381),** 1–12.

Burwen, D. R., La Voie, L., Braun, M. M., Houck, P., Hudson, R., and Ball, R. (2006). Evaluating adverse events after vaccination in the medicare population. *Pharmacoepidemiology and Drug Safety* **15**, 168.

Cameron, J. C., Walsh, D., Finlayson, A. R., and Boyd, J. H. (2006). Oral polio vaccine

and intussusception: A data linkage study using records for vaccination and hospitalization. *American Journal of Epidemiology* **163**, 528–533.

Carlin, J. B., Macartney, K. K., Lee, K. J., et al. (2013). Intussusception risk and disease prevention associated with rotavirus vaccines in Australia's national immunization program. *Clinical Infectious Diseases* **57**, 1427–1434.

Curry, H.B., and Schoenberg, I.J. (1947). On Polya frequency IV: The spline functions and their limits. *Bull. Amer. Math. Soc.* **53,** 1114.

Curry, H.B., and Schoenberg, I.J. (1966). On Polya frequency IV: The fundamental spline functions and their limits.. *J. d'Analuse Math.* **17,** 71-107.

de Boor, C. (1978). *A Practical Guide to Splines.* New York: Springer-Verlag.

Delaney, J. A., and Suissa, S. (2009). The case-crossover study design in pharmacoepidemiology. *Statistical Methods in Medical Research* **18**, 53–65.

Dimatteo, I., Genovese, C.R. and Kass, R.E. (2001). Bayesian curvefitting with free-knot splines. *Biometrika* **88**, 1055–1071.

Dodd, C. N., Romio, S. A., Black, S., et al. (2013). International collaboration to assess the risk of Guillain-Barre' Syndrome following Influenza A (H1N1) 2009 monovalent vaccines. *Vaccine* **31**, 4448-4458.

Douglas, I.J., Evans, S. J., Pocock, S., and Smeet, L. (2009). The risk of fractures associated with thiazolidinediones: A self-controlled sase-series Study. *PLoS Medicine* **6(9)**, e1000154.

Dourado, I., Cunha, S., Teixeira, M. D., et al. (2000). Outbreak of aseptic meningitis associated with mass vaccination with a Urabe-containing measles-mumps-rubella vaccine - Implications for immunization programs. *American Journal of Epidemiology* **151**, 524–530.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–102.

Escolano, S., Hill, C., and Tubert-Bitter, P. (2013). A new self-controlled case series method for analyzing spontaneous reports of adverse events after vaccination. *American Journal of Epidemiology* **178**, 1496–1504.

Farez, M. F., Ysrraelit, M. C., Fiol, M., and Correale, J. (2012). H1N1 vaccination does not increase risk of relapse in multiple sclerosis: A self-controlled case-series study. *Multiple Sclerosis* **18**, 254–256.

Farrington, C. P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* **51**, 228–235.

Farrington, C. P. (2004). Control without separate controls: evaluation of vaccine safety using case-only methods. *Vaccine* **22**, 2064–2070.

Farrington, C.P., Anaya-Izquierdo, K., Whitaker, H. J., Hocine, M. N., Douglas, I., and Smeeth, L. (2011). Self-controlled case series analysis with event-dependent observation periods. *Journal of the American Statistical Association* **106**, 417–426.

Farrington, C. P., Hocine, M.N. (2010). Within-individual dependence in self-controlled case series models for recurrent events. *Journal of the Royal Statistical Society Series C* **59**, 457-475.

Farrington, C. P., Miller, E., and Taylor, B. (2001). MMR and autism: further evidence against a causal association. *Vaccine* **19**, 3632–3635.

Farrington, C. P., Nash, J., and Miller, E. (1996). Case series analysis of adverse reactions to vaccines: A comparative evaluation. *American Journal of Epidemiology* **143**, 1165–1173.

Farrington, P., Pugh, S., Colville, A., Flower, A., Nash, J., Morgan-Capner, P., Rush, M., and Miller, E. (1995). A new method for active surveillance of adverse events from diphteria-tetanus-pertussis and measles mumps rubella vaccines. *Lancet* **345**: 567–569.

Farrington, C. P., and Whitaker, H. J. (2006). Semiparametric analysis of case series data (with Discussion). *Journal of the Royal Statistical Society Series C-Applied Statistics* **55,** 553–580.

Farrington, C.P., Whitaker, H. J., and Hocine, M. N. (2009). Case series analysis for censored, perturbed or curtailed post-event exposures. *Biostatistics* **10,** 3–16.

France, E. K., Glanz, J.M., Xu, S, Davis, R.L., et al.(2004). Safety of the trivalent inactivated influenza vaccine among children. *Archives of Pediatric and Adolescent Medicine* **158**, 1031–1036.

France, E. K., Glanz, J., Xu, S., et al. (2008). Risk of immune thrombocytopenic purpura after measles-mumps-rubella immunization in children. *Pediatrics* **121**, E687–E692.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics* **19,** 1–67.

Galeotti, F., Massari, M., D'Alessandro, R., et al. (2013). Risk of Guillain-Barre' syn-

drome after 2010-2011 influenza vaccination. *European Journal of Epidemiology* **28**, 433–444.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. (2014a). Self controlled case series method with smooth age effect. *Statistics in Medicine* **33(4)**, 639 – 649.

Ghebremichael-Weldeselassie, Y., Whitaker, H. J., and Farrington, C. P. (2014b). Flexible modelling of vaccine effect in self-controlled case series models. *Submitted*

Glanz, J. M., McClure, D.L., Xu, S, Hambidge, S.J., et al. (2006). Four different study designs to evaluate vaccine safety were equally validated with contrasting limitations. *Journal of Clinical Epidemiology* **59**, 808–818.

Gold, M., Dugdale, S., Woodman, R. J., and McCaul, K. A. (2010). Use of the Australian Childhood Immunisation Register for vaccine safety data linkage. *Vaccine* **28**, 4308–4311.

Green, P.G. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman & Hall.

Greenland, S. (1995a). Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology* **6(4)**, 450–454.

Greenland, S. (1995b). Dose-response and trend analysis in epidemiology - Alternatives to categorical analysis. *Epidemiology* **6(4)**, 356–365.

Grosso, A., Douglas, I., MacAllister R., Petersen, I., Smeeth, L., and Hingorani, A. D. (2011). Use of the self-controlled case series method in drug safety assessment. *Expert Opinion Drug Safety* **10(3)**, 337–340.

Gwini, S. M., Coupland, C. A. C., and Siriwardena, A. N. (2011). The effect of influenza vaccination on risk of acute myocardial infarction: Self-controlled case-series study. *Vaccine* **29**, 1145–1149.

Hambidge, S. J., Glanz, J. M., France, E. K., et al. (2006). Safety of trivalent inactivated influenza vaccine in children 6 to 23 months old. *Jama-Journal of the American Medical Association* **296**, 1990–1997.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* New York: Springer-Verlag.

Hauptmann, M., Wellmann, J., Lubin J.H., Rosenberg, P.S., and Kreienbrock, L. (2000). Analysis of exposure-time-response relationships using a spline weight function. *Biometrics* **56(4),** 1105–1108.

Hauptmann, M., Berhane, K., Langholz, B., and Lubin J.H. (2001). Using splines to analyse latency in the Colorado Plateau uranium miners cohort. *Journal of Epidemiology and Biostatistics* **6,** 417–424.

Hauptmann, M., Pohlabeln, H., Lubin, J.H., Jckel, K.H., Ahrens, W., Brske-Hohlfeld, I., and Wichmann, H.E. (2002). The exposure-time-response-relationship between occupational asbestos exposure and lung cancer in two German case-control studies. *American Journal of Industrial Medicine* **41,** 89–97.

Hens,N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P., and Beutels, P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data.* New York: Springer.

Hocine, M. N., Farrington, C. P., Touze, E., et al. (2007). Hepatitis B vaccination and first central nervous system demyelinating events: Reanalysis of a case-control study using the self-controlled case series method. *Vaccine* **25**, 5938–5943

Hocine, M. N., Musonda, P., Andrews, N. J., and Paddy Farrington, C. (2009). Sequential case series analysis for pharmacovigilance. *Journal of the Royal Statistical Society Series A-Statistics in Society* **172**, 213–236.

Hocine, M. Guillemot, D, Tubert-Bitter, P, Moreau, T. (2005). Testing independence between two Poisson-generated multinomial variables in case-series and cohort studies. *Statistics in Medicine* **24**, 4035–4044.

Huang, W.T., Gargiullo, P.M., Broder, K.R., Weintraub, E.S., Iskander, J.K., Klein, N.P., and Baggs, J.M. (2010). Lack of Association Between Acellular Pertussis Vaccine and Seizures in Early Childhood. *Lancet* **126,** E263–E269.

Hughes, R. A., Charlton, J., Latinovic, R., and Gulliford, M. C. (2006). No association between immunization and Guillain-Barre' syndrome in the United Kingdom, 1992 to 2000. *Archives of Internal Medicine* **166**, 1301–1304.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics* **54,** 185–194.

Joly, P., and Commenges, D.(1999). A Penalized Likelihood Approach for a Progressive Three-State Model with Censored and Truncated Data: Application to AIDS. *Biometrics* **55(3),** 887–890.

Joly, P., Commenges, D., Helmer C., and Letenneur L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* **3(3),** 433–443.

Juurlink, D. N., Stukel, T. A., Kwong, J., et al. (2006). Guillain-Barre' syndrome after influenza vaccination in adults - A population-based study. *Archives of Internal Medicine* **166**, 2217–2221.

Klein, N. P., Fireman, B., Yih, W.K., et al. (2010). Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. *Pediatrics* **126**, e1-e8.

Kramarz, P., DeStefano, F., Gargiullo, P. M., et al. (2000). Does influenza vaccination exacerbate asthma? Analysis of a large cohort of children with asthma. *Archives of Family Medicine* **9**, 617–623.

Kramarz, P., DeStefano, F., Gargiullo, P. M., et al. (2001). Does influenza vaccination prevent asthma exacerbations in children? *Journal of Pediatrics* **138**, 306–310.

Kuhnert, R., Hecker, H., Poethko-Mller, C., et al. (2011). A modified self-controlled case series method to examine association between multidose vaccinations and death. *Statistics in Medicine* **30**, 666–677.

Kuhnert, R., Schlaud, M., Poethko-Mller, C., et al. (2012). Reanalyses of case-control studies examining the temporal association between sudden infant death syndrome and vaccination. *Vaccine* **30**, 2349–2356.

Langholz, B., Thomas, D., Xiang, A., and Stram, D. (1999). Latency Analysis in Epidemiologic Studies of Occupational Exposures: Application to the Colorado Plateau Uranium Miners Cohort. *American Journal of Industrial Medicine* **35,** 246–256.

Lee, K.J., and Carlin, J.B. (2014). Fractional polynomial adjustment for time-varying covariates in a self-controlled case series analysis. *Statistics in Medicine* **33(1)**, 105-116.

Lumley, T., Levy, D. (2000). Bias in the case-crossover design: implications for studies of air pollution. *Environmetrics* **11**, 689-704.

Miller, D.L., Ross, E.M., Alderslade, R., Bellman, M.H., and Rawson, N.S.B. (1981). Pertussis immunization and serious acute neurological illness in children. *British Medical Journal* **282**: 1595–1599.

Miller, E., Andrews, N., Waight, P., and Taylor, B. (2003). Bacterial infections, immune overload, and MMR vaccine. *Archives of Disease in Childhood* **88**, 222–223.

Miller, E., Andrews, N., Grant, A., Stowe, J., and Taylor, B. (2005). No evidence of an association between MMR vaccine and gait disturbance. *Archives of Disease in Childhood* **90**, 292–296.

Miller, E., Andrews, N., Stowe, J., Grant, A., Waight, P., and Taylor, B. (2007). Risks of convulsion and aseptic meningitis following measles-mumps-rubella vaccination in the United Kingdom. *American Journal of Epidemiology* **165**, 704–709.

Miller, E., Waight, P., Farrington, P., Andrews, N., Stowe, J., and Taylor, B. (2001). Idiopathic thrombocytopenic purpura and MMR vaccine. *Archives of Disease in Childhood* **84**, 227–229.

Mohammed, S. M., Sentrk, D., Dalrymple, L. S., and Nguyen, D. V. (2012) Measurement error case series models with application to infection-cardiovascular risk in older patients on dialysis. *Journal of the American Statistical Association* **107**, 1310–1323.

Mohammed, S. M., Dalrymple, L. S., Sentrk, D., and Nguyen, D. V. (2013) Design considerations for case series models with exposure onset measurement error. *Statistics in Medicine* **32**, 772–786.

Mullooly, J. P., Pearson, J., Drew, L., et al. (2002). Wheezing lower respiratory disease and vaccination of full-term infants. *Pharmacoepidemiology and Drug Safety* **11**, 21–30.

Mullooly, J. P., Schuler, R., Mesa, J., Drew, L., and DeStefano, F. (2011). Wheezing lower respiratory disease and vaccination of premature infants. *Vaccine* **29**, 7611–7617.

Murphy, T. V., Gargiullo, P. M., Massoudi, M. S., et al. (2001). Intussusception among infants given an oral rotavirus vaccine. *New England Journal of Medicine* **344**, 564–572.

Musonda, P., Farrington, C. P., and Whitaker, H. J. (2006). Sample sizes for self-controlled case series studies. *Statistics in Medicine* **25**, 2618–2631.

Musonda, P., Hocine, M.N., Whitaker, H.J., Farrington, C.P. (2008a). Self-controlled case series analyses: small-sample performance. *Computational Statistics and Data Analysis* **52**, 1942–1957.

Musonda, P., Hocine, M.N., Andrews, N.J., Tubert-Bitter, P., Farrington, C.P. (2008b). Monitoring vaccine safety using case series cumulative sum charts. *Vaccine* **26**, 5358-5367.

Mutsch, M., Zhou, W. G., Rhodes, P., et al. (2004). Use of the inactivated intranasal influenza vaccine and the risk of Bell's palsy in Switzerland. *New England Journal of Medicine* **350**, 896–903.

Naleway, A. L., Belongia, E. A., Donahue, J. G., Kieke, B. A., Glanz, J. M., and Vaccine

Safety, D. (2009). Risk of immune hemolytic anemia in children following immunization. *Vaccine* **27**, 7394–7397.

Navidi, W. (1998). Bidirectional case-crossover designs for exposures with time trends. *Biometrics* **54** , 596-605.

NHS (2013). http://www.nhs.uk/conditions/diabetes-type2/pages/introduction.aspx *Accessed 23/06/2013.*

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statist. Sci.* **1**, 505–527.

O'Sullivan, F. (1988a). Fast computation of fully automated log-density and log-hazard estimators. *Siam Journal on Scientific and Statistical Computing* **9(2)**, 363–379.

O'Sullivan, F. (1988b). Non parametric estimation of relative risk using splines and cross-validation. *Siam Journal on Scientific and Statistical Computing* **9(3)**, 531–542.

Payne, D. C., Aranas, A., McNeil, M. M., Duderstadt, S., and Rose, C. E. (2007). Concurrent vaccinations and US military hospitalizations. *Annals of Epidemiology* **17**, 697–703.

Petersen, I., Gilbert, R., Evans, S., Ridolfi, A., and Nazareth, I. (2010). Oral antibiotic prescribing during pregnancy in primary care: UK population-based study. *Journal of Antimicrobial Chemotherapy* **65**, 2238–2246.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science* **3,** 425–461.

Ramsay, J. O. and Silverman, B.W. (1997). *Functional Data Analysis.* New York: Springer-Verlag.

Romio, S., Weibel, D., Dieleman, J. P., et al. (2014). Guillain-Barre' Syndrome and Adjuvanted Pandemic Influenza A (H1N1) 2009 Vaccines: A Multinational Self-Controlled Case Series in Europe. *PLoS ONE* **9.**

Rondeau, V., and Gonzalez, J.R. (2005). Frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer Methods and Programs in Biomedicine.* **80,** 154–164.

Royston, P., and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Appl. Statist..* **43,** 429–467.

Royston, P., and Altman, D.G. (1997). Approximating statistical functions by using fractional polynomial regression. *Computer Methods and Programs in Biomedicine.* **46(3),** 411–422.

Ruppert, D. Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Sardinas, M. A. G., Cardenas, A. Z., Marie, G. C., et al. (2001). Lack of association between intussusception and oral polio vaccine in Cuban children. *European Journal of Epidemiology* **17**, 783–787.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. R. Statist. Soc. B* **47,** 1–52.

Simpson, S.E. (2013). A Positive Event Dependence Model for Self-Controlled Case Series with Applications in Postmarketing Surveillance. *Biometrics* **69,** 128–136.

Simpson, S. E., Madigan, D., Zorych, I., Schuemie, M. J., Ryan, P. B., and Suchard, M. A. (2013). Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics* **69**, 893–902.

Smeeth, L., Thomas, S. L., Hall, A. J., Hubbard, R., Farrington, P., and Vallance, P. (2004). Risk of myocardial infarction and stroke after acute infection or vaccination. *New England Journal of Medicine* **351**, 2611–2618.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317-344.

Smith, P.L. (1979). Splines as a useful and convenient statistical tool. *American Statistician* **33(2),** 57–62.

Stowe, J., Kafatos, G., Andrews, N., and Miller, E. (2001). Idiopathic thrombocytopenic purpura and the second dose of MMR. *Archives of Disease in Childhood* **93**, 182-183.

Stowe, J., Andrews, N., Wise, L., and Miller, E. (2006). Bell's palsy and parenteral inactivated influenza vaccine. *Human Vaccines* **2**, 110–112.

Stowe, J., Andrews, N., Wise, L., and Miller, E. (2008). Investigation of the Temporal Association of Guillain-Barre' Syndrome With Influenza Vaccine and Influenzalike Illness Using the United Kingdom General Practice Research Database. *American Journal of Epidemiology* **169**, 382-388.

Stowe, J., Andrews, N., Taylor, B., and Miller, E. (2009). No evidence of an increase of bacterial and viral infections following Measles, Mumps and Rubella vaccine. *Vaccine* **27**, 1422–1425.

Suissa, S. (2007). Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology* **167**, 492–499.

Sylvestre, J., and Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine* **28(27),** 3437–3453.

Tata, L. J., West, J., Harrison, T., Farrington, P., Smith, C., and Hubbard, R. (2003). Does influenza vaccination increase consultations, corticosteroid prescriptions, or exacerbations in subjects with asthma or chronic obstructive pulmonary disease? *Thorax* **58**, 835–839.

Taylor, B., Miller, E., Farrington, C. P., et al. (1999). Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *Lancet* **353**, 2026-2029.

Taylor, B., Andrews, N., Stowe, J., Hamidi-Manesh, L., and Miller, E. (2007). No increased risk of relapse after meningococcal C conjugate vaccine in nephrotic syndrome. *Archives of Disease in Childhood* **92**, 887–889.

Thomas, D.C. (1988). Models for exposure-time-response relationships with applications to cancer epidemiology. *Annual Reviews of Public Health* **9,** 451–482.

Traversa, G., Spila-Alegiani, S., Bianchi, C., et al. (2011). Sudden unexpected deaths and vaccinations during the first two years of life in Italy: A case series study. *PLoS ONE* **6**, e16363.

Vacek, P.M. (1997). Assessing the effect of intensity when exposure varies over time. *Statistics in Medicine* **16,** 505–513.

van der Maas N.A.T, Bondt, P.E.V.-d., de Melker, H., Kemmeren, J.M. (2009). Acute cerebellar ataxia in the Netherlands: A study on the association with vaccinations and varicella zoster infection. *Vaccine* **27**: 1970–1973.

Vines, S. K., Farrington, C. P. (2001). Within-subject exposure dependency in case-crossover studies. *Statistics in Medicine* **20**, 3039–3049.

Wahba, G. (1983). Bayesian confidence intervals for the cross validated smoothing spline. *J. R. Statist. Soc. B* **45,** 133-150.

Wand, M.P., Jones M.C. (1995). *Kernel Smoothing.* London: Chapman & Hall/CRC.

Ward, K. N., Bryant, N. J., Andrews, N. J., et al. (2007). Risk of serious neurologic disease after immunization of young children in Britain and Ireland. *Pediatrics* **120**, 314–321.

Wasserman, L. (2006). *All of Nonparametric Statistics.* New York: Springer.

Weinberg, C.R. (1995). How bad is categorization. *Epidemiology* **6(4),** 345–347.

Weldeselassie, Y. G., Whitaker, H. J., and Farrington, C. P. (2011) Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. *Epidemiology and Infection* **139,** 1805–1817.

Whitaker, H. J., Farrington, C. P., Spiessens, B., and Musonda, P. (2006). Tutorial in biostatistics: The self-controlled case series method. *Statistics in Medicine* **25,** 1768–1797.

Whitaker, H. J., Hocine, M. N., Farrington, C. P. (2007) On casecrossover methods for environmental time series data. *Environmentrics* **18**, 157-171.

Whitaker, H. J., Hocine, M. N., and Farrington, C. P. (2009). The methodology of self-controlled case series studies. *Statistical Methods in Medical Research* **18,** 7–26.

Wood, S. (2006a). *Generalized Additive Models.* Boca Raton: Chapman & Hall/CRC.

Wood, S.N. (2006b). On Confidence intervals for generalized additive models based on penalized regression splines. *Aust. N. Z. J. Stat* **48(4),** 445–464.

Xu, S., Hambidge, S. J., McClure, D. L., Daley, M. F., and Glanz, J. M. (2013). A scan statistic for identifying optimal risk windows in vaccine safety studies using self-controlled case series design. *Statistics in Medicine* **32**, 3290–3299.

Zhao, L. P., and Kolonel, L. N. (1992). Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *American Journal of Epidemiology* **136(4),** 464–474.

Zinman, L., Thoma, J., Kwong, J. C., Kopp, A., Stukel, T. A., and Juurlink, D. N. (2009). Safety of influenza vaccination in patients with myasthenia gravis: A population-based study. *Muscle & Nerve* **40**, 947–951.

Appendix A

Review of Vaccine Studies

This appendix presents the form which was used to review SCCS vaccine studies in Chapter 2.

Type of paper:

Focused on estimating relative incidences for one or more vaccine/adverse event combinations using the case series method (alone or alongside other methods)?	☐
Methodological paper with an example data set? If so give reference for original data but continue filling in form:	☐
Methodological paper with no relevant data on vaccines? (If so, stop now).	☐
General paper (eg review, or epidemiology paper) with only passing reference to case series methods? (If so, stop now).	☐

Vaccines and adverse events studied (If there are several, list just the main result or results, and indicate there are others):

Vaccine	Adverse event	Post vaccination risk period	RI (CI)

Data on events and vaccination:

Clear description provided of how data were obtained?

Sufficient detail to verify that ascertainment of vaccinations and events were independent?

Precise dates available, or imputed (if the latter, give details)?

Repeat events excluded or included (if included, give detail of how separate episodes are defined)?

Vaccines given in single or multiple doses (give details)?

Was a case note review undertaken (give details: all or sample)?

Study type:

Hypothesis generating (no prior hypothesis)?

Confirmatory (first study, but based on a prior hypothesis)?

Repeat (previous studies already undertaken)?

Not clear which of the above?

Did study involve a comparison of case series with another method (give details)?

Population and observation period:

Age range of cases:

Calendar period of study:

Observation period rigorously defined (in such a way that analysis could be repeated)?

Age groups (and other temporal adjustments) used in analysis:

Specified rigorously (in such a way that analysis could be repeated)?

Specified vaguely?

Used but not specified?

Give details of age groups: how many, how wide?

Sensitivity to age groupings investigated?

No age stratification used?

Any other temporal adjustment(season, year etc)?

Give details

Rationale for risk periods used in analysis:

Was the choice of risk period(s) based on prior studies?

Based on general knowledge, but not previous studies?

Not justified in any way?

Exogeneity assumption:

Did the authors discuss whether the assumption is likely to hold, namely (a) observation periods do not depend on event (b) events do not affect exposures

Was a pre-exposure risk period used (if so, give details)?

Other relevant discussion or methods used?

Sample size details:

Number of cases and events included in the analysis (for analyses of several events or vaccines, give full details).

Other relevant details:

Any relevant exclusions or inclusions?

Other statistical features:

Tests for interaction with fixed covariates?

Dose-specific effects investigated?

p-values quoted for vaccine effects?

Software used?

Anything else of interest?

Other comments:

Unusual features:

Good practice:

Bad practice:

Details of results with other methods (if used):

Any other comments: